AMERICAN UNIVERSITY OF BEIRUT

EMPLOYEE TURNOVER PREDICTION WITH MACHINE LEARNING ALGORITHMS

by
MELISSA JOSEPH LOUAK

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Industrial Engineering and Management
of the Maroon Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
May, 2021

AMERICAN UNIVERSITY OF BEIRUT

EMPLOYEE TURNOVER PREDICTION WITH MACHINE LEARNING ALGORITHMS

by

MELISSA JOSEPH LOUAK

Approved by:

J Azar

[Dr. Jimmy Azar, Assistant Professor] Advisor
[Industrial Engineering and Management]

Nadine Marie Moacdieh

[Dr. Nadine Moacdieh, Assistant Professor] Member of Committee
[Industrial Engineering and Management]

[Dr. Maher Nouiehed, Assistant Professor] Member of Committee
[Industrial Engineering and Management]

Date of thesis defense: May 3, 2021

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: _____Louak_____Melissa_____ Joseph _____
Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

☐ As of the date of submission

☒ One year from the date of submission of my thesis.

☐ Two years from the date of submission of my thesis.

☐ Three years from the date of submission of my thesis.

08/05/2021

Signature Date

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my thesis supervisor Dr. Jimmy Azar, who was following up, guiding me, and sharing with me his strong technical background in machine learning and data analytics. He was very supportive throughout the thesis preparation and very meticulous to details. I am sincerely proud that I was under the supervision of Dr. Jimmy Azar. I learned a lot from his continuous support and mentoring throughout this period.

I would like to thank the thesis committee members, Dr. Nadine Moacdieh and Dr. Maher Nouiehed, for being part of the committee and sharing their insightful experiences and feedback.

Lastly, I would like to thank my family and my friends for their continuous support and encouragement.

ABSTRACT
OF THE THESIS OF

Melissa Joseph Louak for Master of Engineering
Major: Engineering Management

Title: Employee turnover prediction with machine learning algorithms

Background. Human Resource departments hire employees based on behavioral and technical assessments, but sometimes these decisions can be biased. Firms incur losses in terms of time and hiring costs when an employee resigns. Machine learning algorithms can help alleviate this problem if applied correctly. Such algorithms can process bulk data and automate processes that would have been otherwise slow and tedious.

Objectives. This study aims to explore factors that influence attrition and thus help reduce the cost a company incurs. We use machine learning algorithms to predict employee turnover, analyze the factors that lead to employee attrition, and group them by their impact on the number of years an employee is willing to stay in the company.

Methods. We test and compare several machine learning algorithms applied on our (fictitious) dataset such as Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, Adaptive Boosting, Support Vector Machine, K-nearest neighbors, and Artificial Neural Networks. We also evaluate the features that contribute to attrition by ranking them from the most to least important . Moreover, we build a regression model and highlight the features mostly correlated with employee attrition whether positively or negatively. Finally, we present a retention plan to avoid attrition based on our collective results deduced from the analysis.

Results. Random Forest gave the best results on our dataset in terms of AUROC and other evaluation measures. The most important features that influenced attrition were overtime, total satisfaction score, marital status, stock options level, and monthly income. Moreover, age and monthly income showed a positive correlation with the number of years an employee stayed at the company, whereas the distance from work to home and the number of companies an employee had worked in showed a negative correlation.

Conclusion. The thesis findings highlight the reasons behind employee attrition. We provide detailed recommendations based on our results for reducing attrition and lowering attrition costs. The approach and methodology followed in this work can be elaborated and applied on real-world HR datasets.

TABLE OF CONTENTS

ILLUSTRATIONS

Figure

TABLES

Table

ABBREVIATIONS

ANN	Artificial Neural Network
AUC	Area Under the Curve
CV	Curriculum Vitae
FN	False Negative
FP	False Positive
HR	Human Resource
IBM	International Business Machines
KNN	K-Nearest Neighbors
PCA	Principal Component Analysis
RBF	Radial Basis function
RMSEA	Root Mean Squared Error of Approximation
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-Sampling Technique
TN	True Negative
TP	True Positive
SVM	Support Vector Machine
VOBN	Variable-Order Bayesian Network

CHAPTER I

LITERATURE REVIEW

A. Overview of Employee Attrition

Bill Gates once said, "You take away our top 20 employees and we become a mediocre company" (Yadav, Jain, & Singh, 2018). One of the most general definitions of attrition is the reduction in the number of employees due to resignation, death, or retirement (Nappinnai & Premavathy, 2013). Moreover, according to the Longman Dictionary of Contemporary English, it could be defined as the inability to replace employees when they leave the company. Employee attrition is a crucial factor that affects the company's revenues. A company suffers from a very high cost when an employee leaves, such as additional recruitment cost, low production cost, cost of training, loss of potential customers, and loss of sales (Anand, Saravanasudhan, & Vijesh, 2012). Employee churn or attrition is a serious issue for high-tech companies since it is hard to find the appropriate candidate for jobs that require special skills and experiences. Secondly, recruiting an employee that fits a position requires a lot of time and money. Thirdly, the loss of an employee could negatively affect customer satisfaction. According to Abbott, employee morale and customer satisfaction are strongly correlated. Satisfied employees are willing to provide good services and products. Therefore, their happiness at work helps them earn customer loyalty and contentment (Abbott, 2003). In their turn, they should learn new expertise to meet the required levels achieved by other employees (Saradhi & Palshikar, 2011). A study suggested that if employees are positive about their job, they can accept any change that happens within the company and most importantly organizational change. It also

proposes increasing employees' salaries will lead to a positive attitude towards the job and improved productivity towards organizational change (Shah, Irani, & Sharif, 2017). Promotions are considered accomplishments and moral satisfaction to employees and it gives them an attachment feeling and a psychological achievement to the company (Judge, Cable, Bourdeau, & Bretz Jr., 1995). The employee's turnover is not only determined by the behavioral factor, but it is also dominated by the attitudinal aspects. It can cause tangible losses such as training and recruitment costs and intangible losses including the risk of lacking the know-how (Yan & Zhou, 2010). Furthermore, the employee's mental health strongly affects their productivity and their will to engage in the organization (Holt, Armenakis, Feild, & Harris, 2007). Employee churn or attrition is divided into two main categories: voluntary and involuntary churn. Involuntary churn is when a company expels the employee because of dissatisfaction with the employee's services. However, voluntary churn is when the employee decides to leave the company because of a better opportunity in another company, a closer location, a higher salary, and/or many other factors (Saradhi & Palshikar, 2011). In our study, we will focus on understanding the features that contribute to voluntary churn.

Employees' churn rate is defined as the ratio of the number of employees resigned at the end of the year over the total number of employees at the beginning of the year (Eric, 2020). However, the period can vary, it can be monthly, quarterly, or yearly. This is the general formula of turnover:

$$Turnover\ rate = \frac{Number\ of\ terminates\ at\ the\ end\ of\ the\ period}{Number\ of\ employees \in the\ beginning\ of\ the\ period}$$

This rate is approximated to 12-15% yearly in the IT or high tech industry which is a high rate. By assuming a low churn rate of 5% the cost of an employee leaving is approximately 1.5 times the annual salary of an employee in the IT industry (Dolatabadi & Keynia, 2017). Furthermore, the company has to replace the resigned employee so it will incur a hiring cost and a salary premium as well (Yadav, Jain, & Singh, 2018). A salary premium is defined as the difference between an employee's salary and the salary of the individual replacing that employee after leaving the company (Singh et al., 2012). Therefore, HR Analytics uses data mining techniques such as classification and predictions to overcome these losses and to prevent employee attrition (Singh et al., 2012). The literature review consists of the algorithms used to predict employees' churn and the reasons why employees leave the company. It will be regarded from two perspectives: descriptive which includes statistics as well and prescriptive.

B. Reasons for attrition

It is important to tackle the root and the main reasons causing employees to leave the company. The reasons behind employee turnover could be positive such as a better job opportunity, salary, or work conditions. It could be a closer work-home location or a better chance for career growth and self-improvement. However, employees might leave the company for negative causes that involve conflicts or misunderstandings with managers and/or colleagues, dissatisfaction at work, low salary, lack of training, and possibility of career growth (Saradhi & Palshikar, 2011). Employee attrition might be voluntary due to personal problems, the location of the workplace, or dissatisfaction at work, or having a heavy workload (Dolatabadi & Keynia, 2017). Employees' willingness to change can be linked to their behavior and attitude towards

the assigned job or the communication between their colleagues. Behavioral characteristics are associated with promotions and salary hikes and attitudinal features are identified as the employees' loyalty and organization towards the job (Shah, Irani, & Sharif, 2017). Employees who had low salaries compared to the average employee salary are more likely to leave their job (Sisodia, Vishwakarma, & Pujahari, 2017). Low income impacts the ability to live and eat properly. Low-paid jobs constitute a major role in employee productivity leading to employee churn. Sisodia et al.'s 2017 study also showed that getting promotions and having a fair workload decrease turnover propensity. Furthermore, hard-working employees having high salaries and not getting any promotions are more likely to leave the company (Sisodia, Vishwakarma, & Pujahari, 2017). It has been validated that the salary has the highest impact on employee contentment. The second factor contributing to job satisfaction is promotions (Shah, Irani, & Sharif, 2017). Thus, economic rewards impact the stability and satisfaction of employees in the organization. A study showed that gender, environmental factor, employee behavior, and educational background are strongly related to attrition. Female workers tend to leave their job more than male workers. Furthermore, GDP growth strongly impacts employee turnovers. One additional factor is education, the higher the education level, the higher the turnover rate for people working in the company for 3 to 5 years (Zhu et al., 2019).

C. Attrition consequences

Employee churn or attrition leads to crucial problems and it is important to understand the motive behind them. The right employee might be hard to find, and this is considered one of the problems of recruitment. Also, the training required for every

new employee has a significant cost for the company (Dolatabadi & Keynia, 2017). Companies incur tangible costs such as hiring costs, salary premium, and intangible costs like productivity losses.

D. Strategies to avoid attrition

Many studies focused on employee turnover through management, psychology, sociology aspects. One of the strategies involves modifying or improving existing policies in the company. Policies include recruitment, selection process, training, salaries, job requirements, and descriptions (Ongori, 2007). An alternative strategy is to implement workforce planning by matching employees' skills and talents with the company's strategy and goals (Rothwell, Alexander, Bernhard, & Books, 2008). Workforce optimization is a strategy that provides good working conditions leading to employee stability and having low turnover rates (Ongori, 2007).

E. Attrition parameters

Common attributes measured in employee attrition are age, tenure, salary, and job satisfaction (Cotton & Tuttle, 1986). In addition to these parameters, other studies added gender, ethnicity, education, marital status, and skill enhancement attrition indicators. A research study (Jain & Nayyar, 2018) suggested 14 elements leading to attrition. Employee turnover is emphasized in the group of employees who were in their 30s. Employees who are 45 years old are most stable in their job. From ages 20 to 40 years old, employees are searching for a satisfying job which they could be stable in. Regarding job satisfaction, a lower job satisfaction level implied a higher attrition rate. Furthermore, 8.1% of employees who were single had left the company, which

represented the highest percentage between different marital status: divorced, married, and single. The divorced employee represented the lowest turnover frequency of 2.3%. The study also showed an inverse proportional relation between the monthly income and the attrition rate suggesting that the high workload with low salaries results in reduced employer satisfaction. New employees who worked for less than a year tend to leave their job which causes problems in the recruitment department. One of the reasons is the miscommunication of the job description. Future employee expectations, which turn out to be different from the job description, cause the newcomer to quit. The job requirements and descriptions should be communicated and explained by the recruiter (Yadav, Jain, & Singh, 2018). The manager has a significant role in maintaining and organizing a team. The attrition rate increases with the replacement of the current manager of a team or a company.

A study by Singh et al. 2012 showed that employees tend to attrite in these cases:

1- If they are in the first few years of their career

2- If they were paid less than 25% below their peers and had a good performance rating, the attrition rate reached 70% (Singh et al., 2012).

Cai et al. show that the average job duration before the turnover is 42 months or about 3 and a half years (Cai et al., 2020).

Customer churn is similar to employee churn but not identical since it is a crucial problem for several companies. The loss of an employee could imply the loss of a customer since the employee is not satisfied with his/her work. Therefore, employee productivity will decrease and the output product or service won't satisfy the customer.

F. Customer churn

Customer churn is the phenomenon when a customer stops using the products/services provided by the company and deviates to other similar products/services from other firms. The loss of a customer not only affects the industry's revenue but also the company's image. Acquiring a new customer is challenging and the loss of a customer affects the company's losses. Similar to employee churn, customer churn is also divided into voluntary and involuntary churn. Voluntary churn refers to the shift of a customer to the services or products of a competitor firm due to customer dissatisfaction with the services provided by the company. An involuntary churn occurs when a company finishes the services of a customer. The main reason behind this termination is the non-payments or late payments of a customer. Customer churn rate can be decreased significantly when implementing several strategies:

1- Credit scoring technique to prevent involuntary churn prediction

2- Product/service quality improvement

3- Product development and innovation

4- Discounts or promotions offerings (Dolatabadi & Keynia, 2017)

5- Customer support service enhancement by minimizing the waiting time (Dolatabadi & Keynia, 2017)

Saradhi and Palshikar 2011 study tackled an employee value model. Some of the customers are profitable for a company. It is essential to characterize the high-value customer to prevent customer churn. The metrics related to evaluating an employee value could be the customer's expected profits in the company. To tackle this problem effectively, strategies should be added to customer churn prediction models. These

strategies should focus on generating methods to detect new customer behavior that includes customer emerging needs and classify customers with distinctive needs (Saradhi & Palshikar, 2011).

G. Descriptive statistics

Several studies consist of statistical analysis to imitate the recruiter's decision. The decision of choosing the best employee is extremely subjective and may be accompanied by numerous judgments and biases that will affect the recruitment process. Therefore, a well-defined objective is established to offer accurate insights to the recruiters. The establishment of an accurate candidate portfolio reduces the recruitment timing procedure by assigning scores to the applicants. Data analysis of employee churn is defined by the descriptive statistics that calculate the mean, standard deviation, and correlation between the variables. Besides, Pearson's correlation determines the linear relationship between the variables. Descriptive statistics can be regarded as one of the preliminary studies. The authors in Shah et al.'s 2017 study endeavored a conceptual model of employee readiness and provide eight hypotheses that lead to attrition. The statistics adopted in this paper (Shah, Irani, & Sharif, 2017) consist of calculating the chi-square, Pearson correlation, and using two forms of analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Shah, Irani, & Sharif, 2017). The hypotheses applied in this study are related to the relationship between salary, job promotion, organizational loyalty, and identification to employee job satisfaction and readiness for change. Another study tested the hypotheses by implementing the structural equation model and comparing the indices of goodness such as the Chi-square and RMSEA that assesses how far is the hypothesized model from the perfect one. The hypotheses focus on the characteristics

leading to employee attrition such as internal and group environment within a particular group and personal features of an employee (age, education background, gender, work experience, etc.(Yan & Zhou, 2010).

H. Previous work

Data mining techniques are effective tools to determine the churn rate and the targeted features for employee churn. Three models were evaluated using K-NN, decision trees, and XGBoost. The latter is a boosting model that is beneficial in regularization to avoid overfitting (Janusz et al., 2018). The model assessment consisted of precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) denoted by AUC. The classification was built using decision trees which serve as a simple and accurate data analytics technique. Precisely, the classification and Regression Trees (CART) algorithm was implemented, which is a recursive partitioning method adopted for dependent and categorical variables (Sivaram & Ramar, 2010).

A comparative study showed that XGBoost proved to be an efficient algorithm to predict employee attrition over glmboost. The accuracy increased from 83% in decision trees to 89% with the XGBoost algorithm. The XGBoost tree algorithm is highly robust and computationally efficient. It can control overfitting since it is based on the principle of gradient boosting (Jain & Nayyar, 2018).

Random forest classifier resulted in the highest accuracy of 98.97% between the KNN, LSVM, Naïve Bayes, and Decision tree classifiers. Also, the random forest has the highest precision, recall, and F-measure (Sisodia, Vishwakarma, & Pujahari, 2017). Due to class imbalance, a study suggested 100 % accuracy for Decision Tree, Naïve Bayes, SVM, and neural network algorithms so the best model was chosen according to

processing speed. Naïve Bayes was the fastest algorithm with 1 second and 81 hundredths of a second (Dolatabadi & Keynia, 2017). VOBN algorithm performance using gradient boosting in the study showed the highest performance while evaluating the area under the curve (AUC = 0.73; Pessach et al., 2020).

Several studies show that the Recursive Feature Elimination increases the precision in the data mining techniques. The implementation of several data mining techniques with and without recursive feature elimination proves that the SVM method has the highest accuracy, precision, and F-measure. The accuracy with feature selection approximately remained the same for Decision Tree, Logistic Regression, KNN, and Random forest methods with feature selection classification. However, the increase in accuracy was very significant for Naïve Bayes (8.2%) and the SVM (4.6%; Yiğit & Shourabizadeh, 2017). Moreover, another study verifies that the random forest has the highest accuracy. Accuracy in the SVM method is very high compared to other algorithms. It shows a high True Positive (resigned/released employees) and True Negative, which can be a severe problem for some companies. By assuming a 20% churn error, the potential churners predicted by the algorithm would be around 28,000 out of 140,000 employees (Saradhi & Palshikar, 2011). Besides, the highest performance in dynamic bipartite graph embedding was through the Random Forest algorithm reaching 89.3% accuracy. To reduce the number of features, PCA was applied for dimensionality reduction. (Cai et al., 2020).

The authors in Yadav et al. 2018 calculated the correlation matrix to find out the most important features that contribute to attrition. They used two approaches to solve the categorical variables problem: the brute-force approach and the one-hot encoding. The first approach consists of dividing the departments into technical and non-technical.

The departments such as sales, HR, marketing, management, and accounting are non-technical, denoted by 0, and the other departments are technical and denoted by 1. The second method consists of transforming the categorical variables clearly by the algorithm. To achieve this, 10 departments were assigned to a one-hot encoding instead of numbers or weights since the Scikit library in python would have given extra value to the feature that has the highest number. Then, recursive feature elimination with cross-validation serves to remove redundant features in the model. The final stage resolves in setting conditions and characteristics that define the experienced employee from a non-experienced one. The conditions considered to evaluate the employee's experience are the time spent in the company, the number of projects, and the last evaluation of the employee. By meeting these conditions, the model was evaluated for experienced employees who have left the company. Moreover, the data mining techniques used in this study are Logistic Regression, SVM, Random Forest, Decision Tree, and AdaBoost. The results significantly improved with the one-hot encoding for the AdaBoost and Random Forest. The classification results of the experienced employees were the highest in the decision tree model (Yadav, Jain, & Singh, 2018). After evaluating different models, Random Forest also proved to be a very effective algorithm in terms of accuracy of 98.97% and precision 99.81%. This high accuracy is due to the imbalanced data and to the low attrition rate in the dataset used. After plotting the receiver operating characteristic (ROC) for different classifiers, the Random Forest also has the highest area under the curve (Sisodia, Vishwakarma, & Pujahari, 2017).

Post-hire prediction could be a suitable performance measure that includes employees' activities and involves an analysis of the real features that contribute to the turnover. However, it may cause late recruitment mistakes due to subjective evaluations

of the candidate. The forecast using pre-hire data has a significant impact on avoiding financial losses. Few studies have predicted the best recruitment strategy and the features that affect the performance from the pre-hire data by evaluating the performance of the candidates (Pessach et al., 2020). The pre-hire approach might be an efficient tool for firm future savings. In Chen-Fu Chien and Li-Fei Chen's 2008 study, the authors used a decision tree, specifically the CHAID algorithm since it is suitable for categorical variables. The evaluation methods consist of prediction accuracy and lift. The lift is a ratio reflecting the change in concentration of a specific class. Therefore, a lift value greater than one shows a robust class (Chien & Chen, 2008). Moreover, categorical data or textual information in the dataset can significantly improve the predictions as suggested by Coussement and Van Den Poel's study (Coussement & Van den Poel, 2008). Many studies predict employee churn using Decision Tree, Naïve Bayes, Logistical Regression, and Support Vector Machines (Huang, Huang, & Kechadi, 2010).

Even though employee turnover analysis is implemented using traditional machine learning methods, deep learning demonstrated significant results when dealing with randomness. One of the studies confirms that the integration of deep learning showed good results with unseen data. The Root Mean Squared Error (RMSE) and the Mean Squared Error (MSE) are the performance measures used for the deep learning model. Once the algorithm converges to an appropriate minimum, the training stops. Umang Soni et al. 2018 study showed that the artificial neural network is much better than the adaptive neuro-fuzzy inference system since it was able to better predict unseen data. In their study, they applied the back-propagation algorithm and logistic sigmoidal function as an activation function to achieve a robust prediction by using multi-layer neural networks. In back-propagation, the error between the actual and predicted value

is calculated and the purpose is to minimize the error and to update the weights accordingly. The training will stop when the error converges to a certain value set by the user. This error should be very small to ignore all types of biases (Soni, Singh, Swami, & Deshwal, 2018).

The essential evaluation metrics are accuracy, precision, recall, and F1 measure. However, when dealing with imbalanced datasets, the best evaluation technique is the ROC and AUC (Chawla, 2005) (Zhu et al., 2019).

The following table represents a summary of some of the major articles of employee attrition.

Table 1-Summary table

Author	Title	Journal	Findings
Janusz et al., 2018	How to Match Jobs and Candidates - A Recruitment Support System Based on Feature Engineering and Advanced Analytics	Springer	XGBoost gave the best result with a precision of 0.163, recall 0.153 F1-score 0.16, and AUC of 0.641.
Jain & Nayyar, 2018	Predicting Employee Attrition using XGBoost Machine Learning Approach	IEEE Xplore	Decision tree reached a 83% accuracy. With XGBoost it reached 89%
Sisodia, Vishwakarma, & Pujahari, 2017	Evaluation of machine learning models for employee churn prediction	IEEE Xplore	Random forest reached an accuracy of 98 % and naïve Bayes the lowest 79 %
Dolatabadi & Keynia, 2017	Designing of customer and employee churn prediction model based on data mining method and neural predictor	ProQuest	SVM: accuracy 84.8 % Random Forest: 79.3 % ANN: 91%
Pessach et al., 2020	Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming	Elsevier	Gradient Boosting AUC = 0.73 Random Forest AUC=0.719 Naïve Bayes AUC = 0.677
Yiğit & Shourabizadeh, 2017	An approach for predicting employee churn by using data mining	IEEE Xplore	With feature selection, SVM is the best algorithm with an accuracy of 0.89, precision 0.89. Then logistic regression with an accuracy of 0.87 and a precision of 0.74
Yadav, Jain, & Singh, 2018	Early Prediction of Employee attrition using data mining techniques	IEEE Xplore	With one hot encoding, Decision tree is the best in terms of

			accuracy (98 %) and precision (0.95)

I. Data cleaning

Data preprocessing or data cleaning is a crucial step before applying any machine learning algorithm. For this reason, missing values can be either replaced by zeros (Pessach et al., 2020) or by their average value. However, replacing a missing value with 0 is not a preferred way of dealing with missing values since the standard deviation between 0 and the other records might be big, leading to the inaccuracy of the model. A better approach is to replace by the mean value.

J. Evaluation metrics

Several evaluation metrics are presented. The confusion matrix is a table that evaluates the performance of the model. It is applied in supervised learning models. The correct predictions are modelled in the True Positive and True Negative labels represented by TP and TN where P and N represents 2 different classes.

Table 2-Confusion matrix

		Predicted
Actual	Turnover	Non-turnover
Turnover	TP	FN
Non-turnover	FP	TN

Accuracy is the ratio of correct predictions over the total predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall refers to the proportion of positive predicted values in the positive samples.

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TP + FN}$$

Precision represents the ratio of positive predictions over the total positive.

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Another evaluation metric to consider is the error which is determined by the cost function, MSE, and RMSE (Soni, Singh, Swami, & Deshwal, 2018).

K. Machine Learning Algorithms

SVM

The goal of the Support Vector Machines to be able to separate the 2 classes through a hyperplane, which represents a decision boundary. The SVM maximizes the distance between the decision boundary and the records to avoid overfitting.

Logistic regression

In logistic regression, the goal of minimizing the loss function is achieved using the gradient descent method by iteratively adjusting the algorithm parameters. The output of the classification is defined as probabilities through the sigmoid function.

Naïve Bayes

Naive Bayes is applied in a supervised model, it computes the posterior probabilities based on the prior probabilities. Given the label variable y and the dependent variables x_1 to x_n, Naïve Bayes computes the probability of y given the dependent variables through the following relationship:

$$P(y|x1, \dots, x_n) = \frac{P(y)P(x1, \dots, x_n|y)}{P(x1, \dots, x_n)}$$

The advantage of this algorithm is the speed. It is fast comparing to other models. This algorithm is fast and performs well with a big dataset. It is a good estimator when dealing with categorical data. However, its accuracy is not high comparing to other sophisticated models due to its simple assumption of the independence between all the variables (Y. Zhang, Yan, & He, 2016). So it assumes that a certain feature x1 is independent of other features such as x2, x3, etc. This algorithm starts by converting the dataset into a frequency table, then creates the probability table, and finally computes the posterior probability table for each category. A major disadvantage of the Naïve Bayes algorithm that it assumes null independence between the variables that's why it is a bad estimator since, in real life, it is hard to have null independence between the attributes. Naïve Bayes has wide applications such as real-

time prediction, recommendation systems, and multi-class classification (Ray et al. 2020).

Decision Tree

A decision tree is a simple algorithm and it is easy to visualize. It consists of branches and nodes. A branch consists of a test on a particular feature and the node denotes a category. In our daily lives, we face many problems and we propose many solutions to reach the most optimal one. Decision trees are based on a decision that is made from different steps. The data is fed into a decision tree. Then, the algorithm divides it into smaller and smaller groups until it reaches a step where the splitting is done. This splitting is done based on the splitting criteria. (Wills, Underwood, & Barrett, 2021). The splitting criteria are the Gini index or Entropy. The Gini formula is represented below, where m represents the terminal node, K is the observations in node m, and p_{mk} represents the probability in this region (Mathan, Kumar, Panchatcharam, Manogaran, & Varadharajan, 2018).

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

Similarly, entropy is computed by taking the logarithmic probability of node m in K observations. However, entropy is computationally expensive.

$$H(Q_m) = -\sum_k p_{mk} log\ (p_{mk})$$

The problem with the decision tree is that it can overfit. For this reason, fine-tuning is recommended such as setting the maximum depth and/or the maximum number of required leaves. Another disadvantage of decision trees is that they are sensitive to any kind of small change and noise. They are very unstable especially with

an imbalanced dataset. That's why it is important to have a balanced dataset before the training (Friedman, Hastie, & Tibshirani, 2001).

Random Forest

Random Forest is a supervised machine learning algorithm. The idea behind the random forest is to choose the class with the highest votes. This algorithm is built based on several decision trees. It combines them and takes the best prediction out of these trees. The advantage of the Random Forest classifier is that not only it serves as a classifier but works as a regressor as well.

KNN

KNN is a simple algorithm that can be applied for classification or regression models. It matches similar features together to determine the class of the test data. Even though this algorithm provides simplicity, a relatively high accuracy, and no assumptions about the data not like the Naïve Bayes algorithm, it is a computationally expensive algorithm and requires a high memory (Bronshtein 2018).

AdaBoost

AdaBoost or Adaptive Boosting is a boosting technique. On each iteration, the weights are adjusted and re-assigned to the next iteration. It is used by combining different weak classifiers to form a strong one and boost the weak ones (Baig, Awais, & El-Alfy, 2017). This algorithm is used in supervised learning to reduce bias and high variance. It works by adjusting the misclassified records from the weak classifier (P.-B.

Zhang & Yang, 2016). It prioritizes the ones that have the highest misclassifications rate. A simple idea of how AdaBoost works is represented in the following pictures.

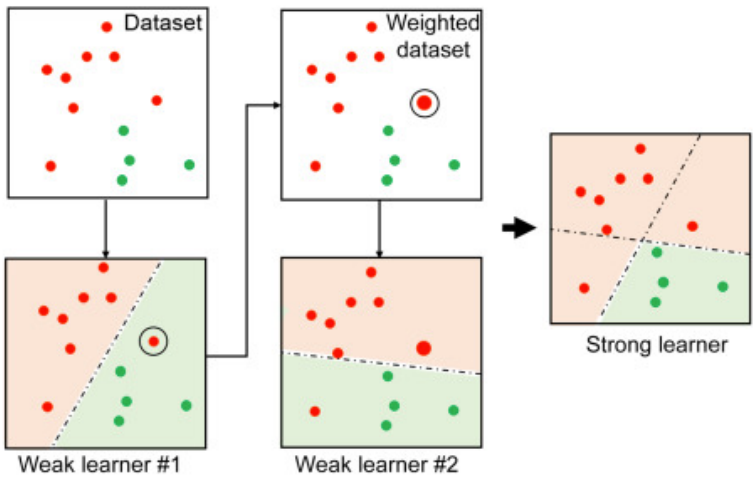

Figure 1-AdaBoost

ANN

Artificial neural networks consist of 3 main layers: input variables, hidden layers, and the output layer. The neural network architecture in a study by Soni consisted of 8 inputs, 1 output, and 2 hidden layers containing 25 neurons in the 1st layer and 50 neurons in the 2nd one. The number of epochs was set to 1000 and the error goal to 0.01. The inputs consist of employee features: satisfaction level, last evaluation, total projects, average monthly hours, number of years spent in the company, work accident, promotion in the last 5 years, and salary. The binary output is employee turnover (Soni, Singh, Swami, & Deshwal, 2018).

XGBoost

The concept of XGBoost works on iteratively producing a new CART. Regularization parameters are set to avoid overfitting.

CHAPTER II

METHODOLOGY

First of all, we will study the correlation between the dependent variable, which is attrition to the other independent variables for a dataset by computing the correlation matrix and plotting histograms. We propose to study the features that affect employees' attrition using Kaggle datasets and to build an employee portfolio that relies on Machine Learning methods such as KNN, Decision Tree, Random Forest, Logistic Regression, SVM, AdaBoost, Naïve Bayes, and deep learning. One of the company's assets is an experienced employee. Consequently, our purpose in this study is to determine the manifold factors that contribute to employee attrition. Data mining techniques proved to be efficient tools for evaluating employee churn. Based on several papers mentioned in the literature review section, choosing the appropriate algorithm leads to high accuracy and precision that can reach up to 98%. The goal is to find the most efficient machine learning algorithm to evaluate the candidate's performance and predict employee churn by testing different classifiers. The aim is to determine key features contributing to accurate predictions of employee churn. In addition to performing binary classification, a regression model will be developed to predict the number of years an employee will remain in a company. Besides highlighting the impact of applying different machine learning algorithms, hyper-parameter tuning will boost the results of our model.
Our performance metrics consist of precision, accuracy, recall, and ROC.

Our approach consists of the following steps:

1- Analyze 4 employee post-hire datasets and choosing the best one by identifying the features applied and their type

2- Clean the data by handling the missing values

3- Compute the correlation matrix and scatter plots to identify the relation between the variables

4- Apply statistics such as Chi-Square distribution to study the dependance between the variables

5- Balance the data

6- Implement different machine learning and deep learning algorithms

7- Evaluate the accuracy, precision, recall, and F-measure metrics on the test set

8- Plot the ROC and evaluate the AUC

9- Select the most relevant features contributing to attrition and grouping them into different groups

10- Build a regression model that predicts how many years an employee will probably stay in the company

11- Evaluate the results by building a retention plan that reflects the most important features contributing to employee turnover

The 9th step consists of grouping the features into two main groups: personal (gender, age, education, marital status) and organizations (promotions, salary, workload, performance, qualifications, industry working in, length of service).

Furthermore, an additional parameter will be included to better understand our data and assess employee's position in the company classifying an employee's salary as low or high based on the difference between an individual salary to the average salary of all employees.

CHAPTER III

DATASETS

The datasets are publicly available on Kaggle. It is a website made for data scientist enthusiasts. Having access to datasets is hard due to the confidentiality of employees' data. After asking access for datasets of several companies in Lebanon and due to inaccessibility of the publicly available datasets, we present the following table containing the publicly available datasets available for HR analytics:

Table 3-Datasets

Dataset	Number of records	Number of Features	Dependent variable
1- IBM	1470	34	Attrition (1/0)
2- HR attrition	54808	14	Promotion (1/0)
3- HR dataset	14999	10	Left
4- HR-IBM dataset	4410	24	Attrition

Even though these datasets are fictional, we chose the IBM dataset to work on since it has the biggest number of features. It is well known and used by most of the data science community. It was created by IBM, which is a multinational company that sells cloud and IT services.

The following table splits the independent variables into categorical nominal and numerical groups. The dependent variable is attrition.

Table 4-Dataset 1

Dataset 1		
Categorical	**Numerical**	
Business Travel	Daily Rate	Age
Department	Distance from home	Percent Salary Hike
Education	Employee Count	Performance Rating
Education field	Employee Number	Standard Hours
Gender	Environment satisfaction	Stock Option Level
Job Involvement	Hourly Rate	Total Working Years
Job role	Monthly Income	Training Times Last Year
Job Satisfaction	Monthly Rate	Work Life Balance
Marital status	Number of Companies	Years At Company
Over 18	worked in	Years in Current Role
Over time		Years since Last
Relationship Satisfaction		Promotion
		Years with Current
		Manager

The following table represents a mapping summary for the levels of some attributes.

Table 5-Dataset 1 attributes levels

Education	**Environment Satisfaction** **Job Involvement** **Job Satisfaction** **Relationship Satisfaction**	**Performance Rating**	**Work Life Balance**
1 : Below College 2: College 3: Bachelor 4: Master 5: Doctor	1: Low 2: Medium 3: High 4: Very High	1: Low 2: Good 3: Excellent 4: Outstanding	1: Bad 2: Good 3: Better 4: Best

CHAPTER IV

RESULTS

A. Descriptive results

Our first main goal is to visualize the data and to analyze it. For this reason, bar plots, box plots, and correlation plots were implemented to analyze and describe the data. First of all, we visualize the feature distributions.

Figure 2-Features Distribution

As we can see, different types of distributions are shown.

Table 6 - Features distributions

Normal Distribution	Right Skewed Distribution	Uniform Distribution
Age	Distance from Home	Daily Rate
Total Satisfaction	Monthly Income	Employee Number
	Number of Companies worked In	Hourly Rate
	Percent Salary Hike	Monthly Rate
	Total Working Years	
	Years at the Company	
	Years in the Current Role	
	Years Since Last Promotion	
	Years with Current Manager	

Since the attributes that have constant values are not changing, we can eliminate them in our study. The monthly income variable can be a substitute for the daily, hourly, and monthly rates.

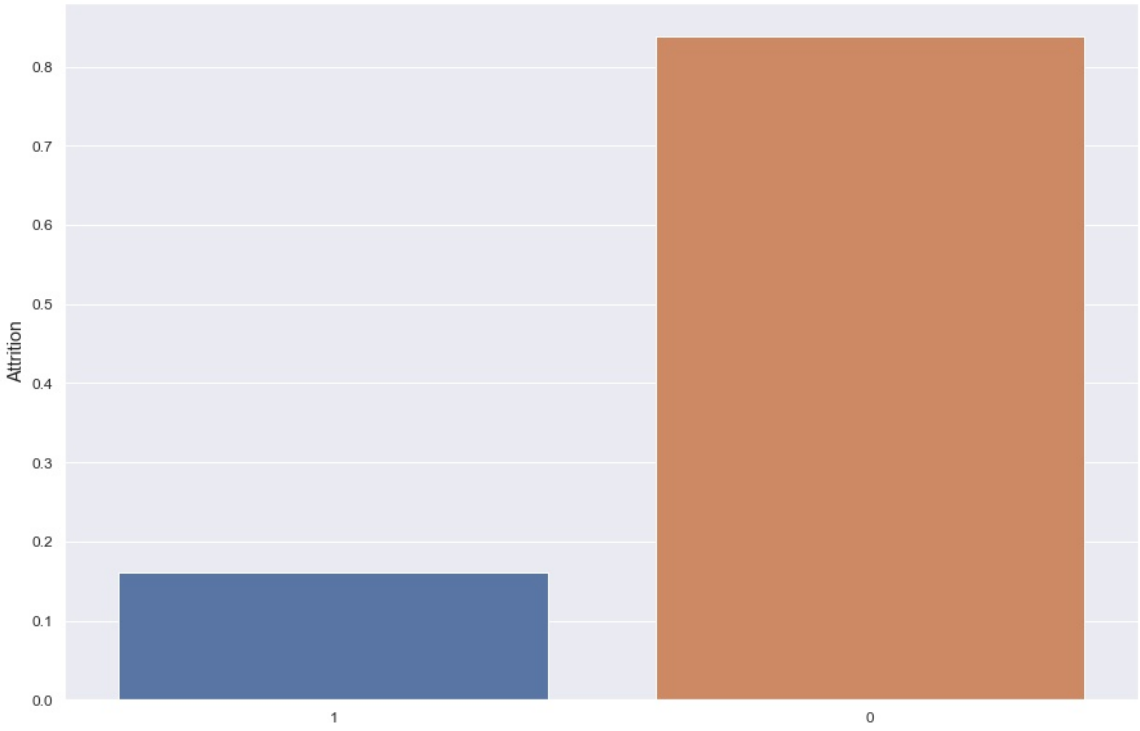

Figure 3-Attrition Percentage

We notice that the attrition rate is equal to 16% and the percentage of employees staying in the company is 84%. We observe the imbalanced data and we propose the SMOTE technique to solve it.

Income Status was an additional parameter implemented to compare the salary of each employee to the average of all the salaries. We can see that the proportion of low salaries is much higher than the high salaries.

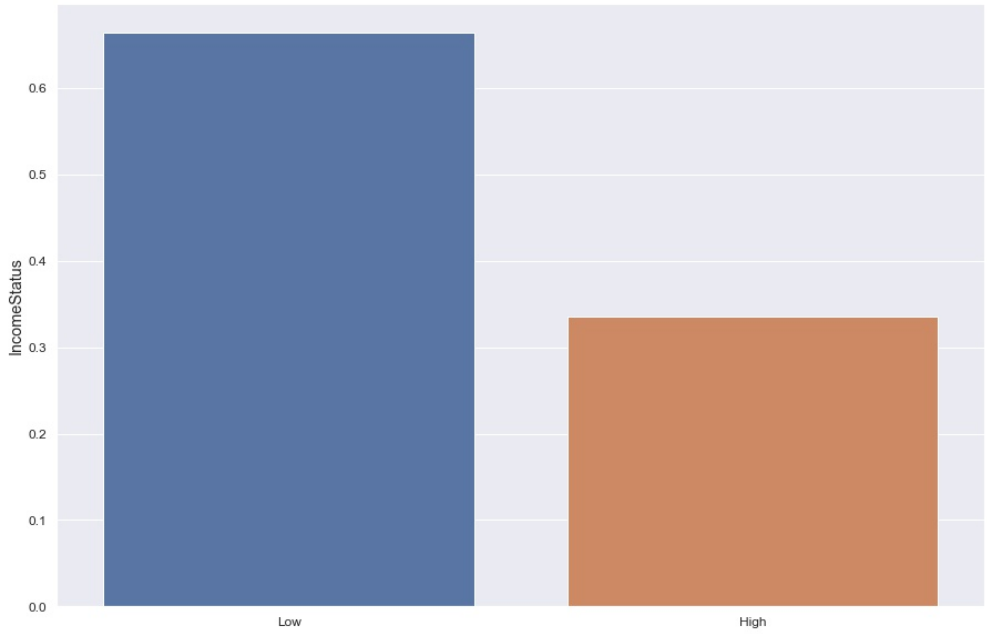

Figure 4-Income Status Percentage

The percentage of employees who have a low-income status (66%) is much higher than the ones who have high-income status (34%).

After doing a correlation plot to find the relationship between different features, we take a lower threshold of 70% for the highly correlated features and we list below the major features from the most correlated to the least ones.

(1) Job level and monthly income (95%)

(2) Job level and total working years (78%)

(3) Performance rating and percent salary hike (77%)

(4) Total working years and monthly income (77%)

(5) Years with current manager and years at the company (77%)

(6) Years in current role and years at the company (76%)

(7) Years with current manager and years in the current role (71%)

Some of the negatively correlated features are listed below by taking an upper threshold of -4%.

All the other features percentages are between -4% and 70%.

(1) Years in current role and number of companies worked in (-9%)

(2) Hourly rate and environment satisfaction (-5%)

(3) Years since last promotion and number of companies worked in (-4%)

(4) Work-life balance and daily rate (-4%)

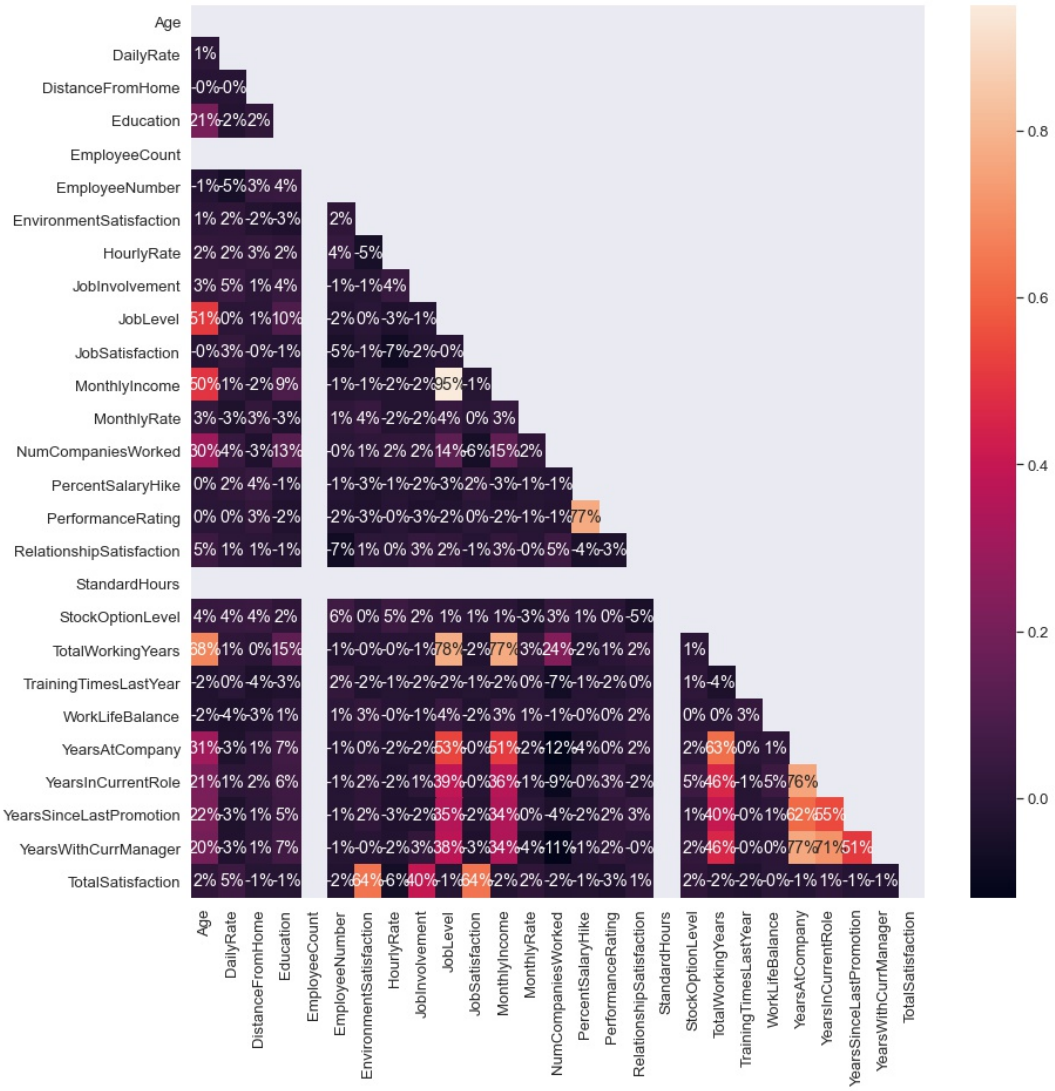

Figure 5-Heatmap

Many attributes were plotted in function of attrition to compare and see the most specific parameter of attrition. The 0s and 1s in the plots represent if the person is an ex-employee (1) or still in the company (0).

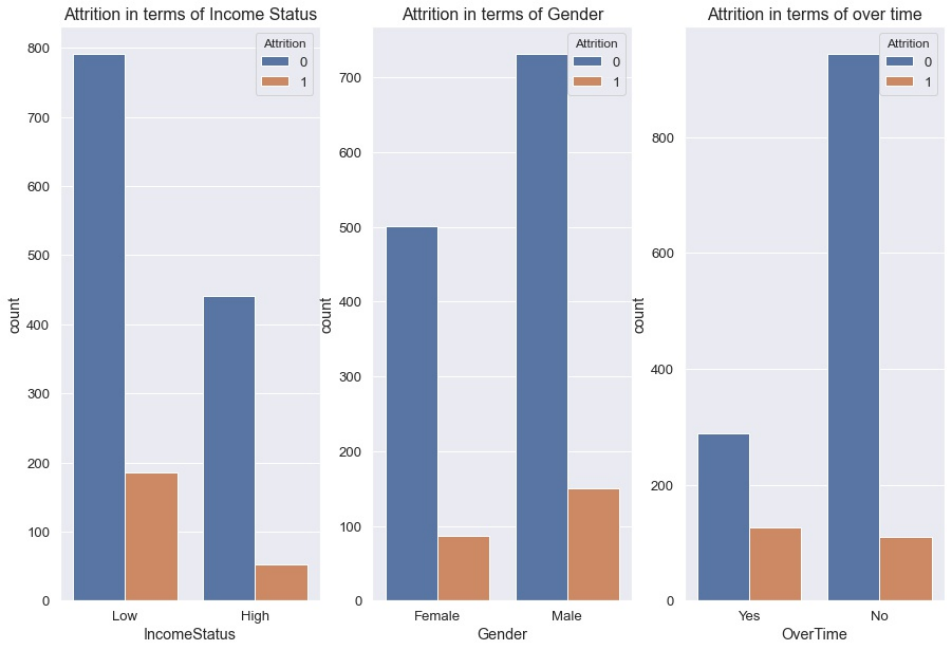

Figure 6-Attrition in terms of income status, gender, and over time

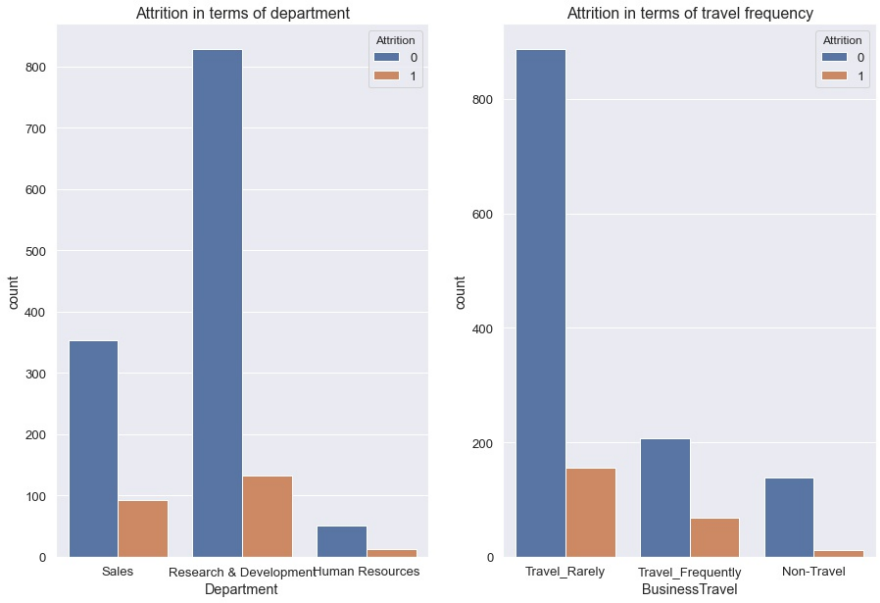

Figure 7-Attrition in terms of department, and business travel frequency

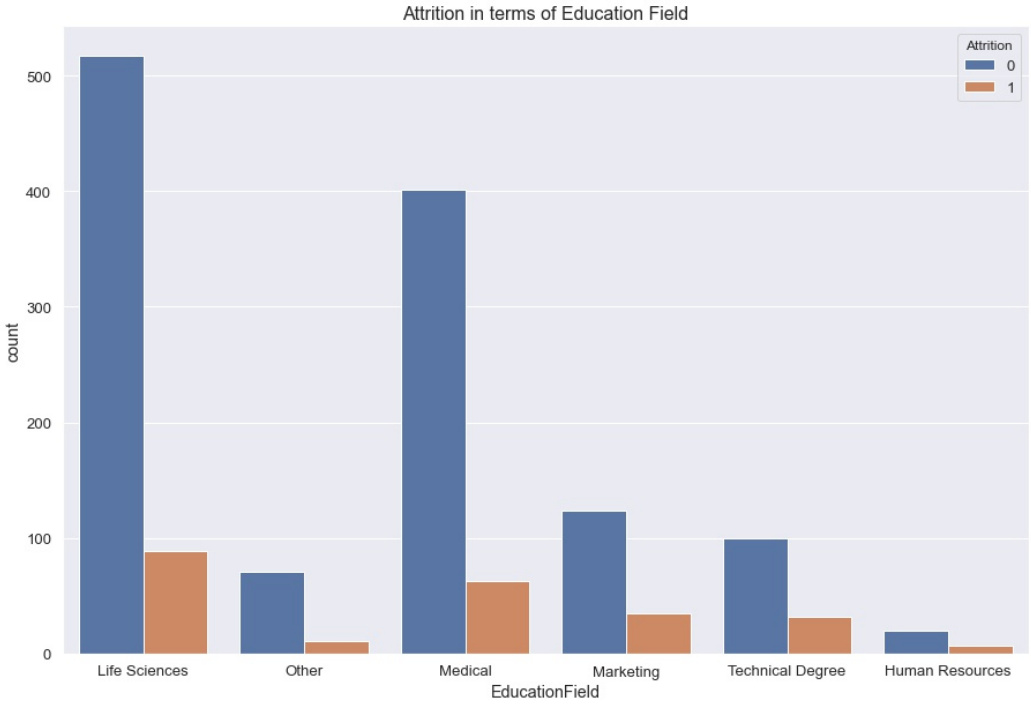

Figure 8- Attrition in terms of education field

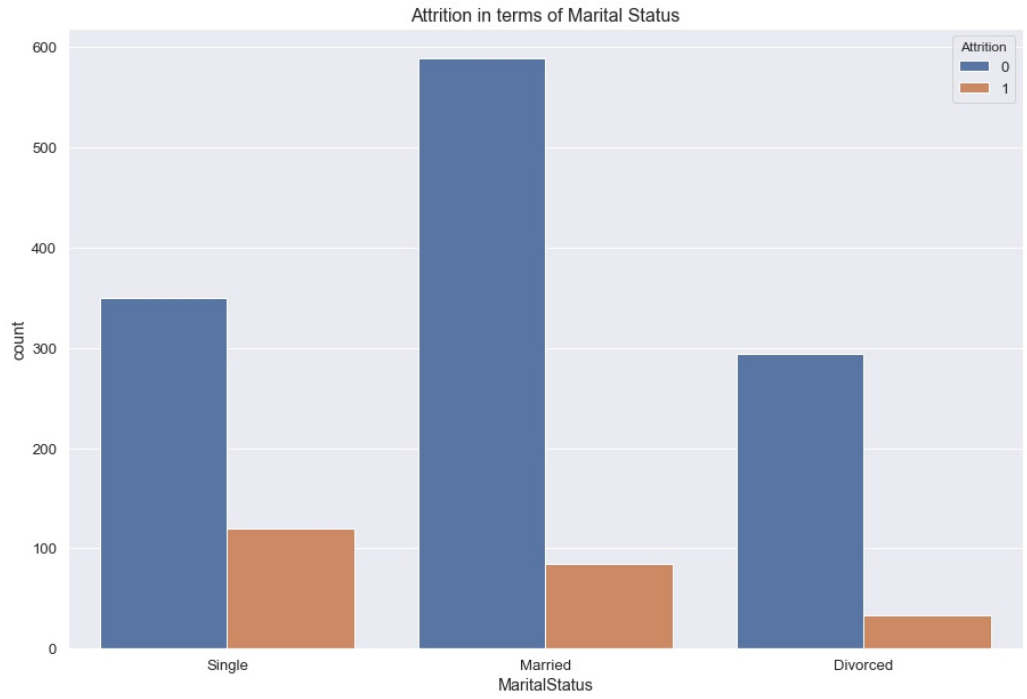

Figure 9-Attrition in terms of marital status

The bar plots give a visual presentation of the proportions of ex-employees and current employees in terms of several factors. However, to visualize the most important category attribute that leads to attrition, we plot the percentage of each category knowing that attrition is equal to one for different attributes.

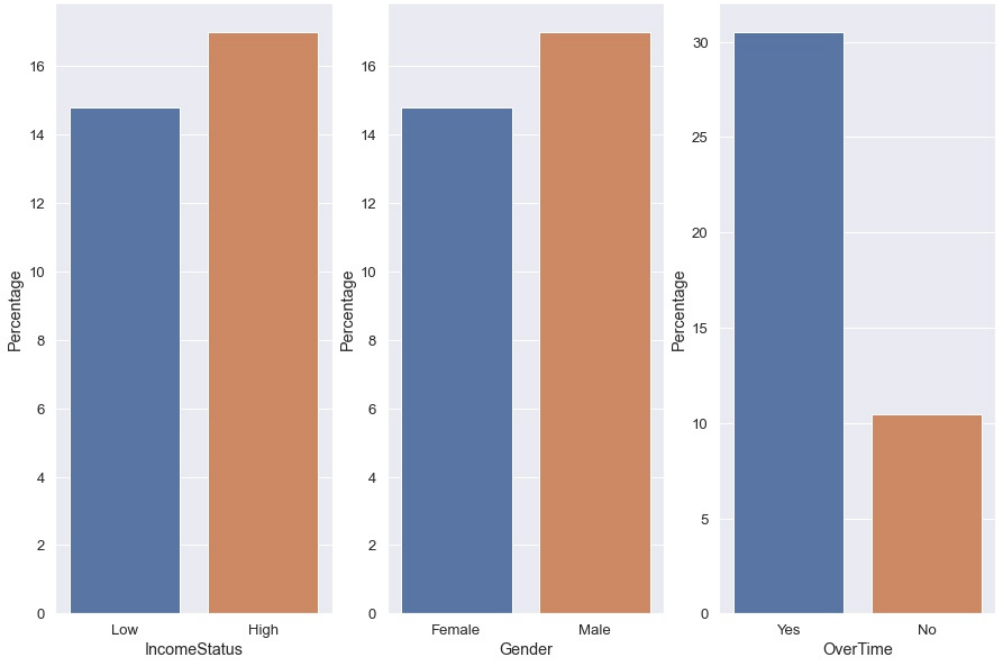

Figure 10-Percentage of leavers in terms of income status, gender, and over time

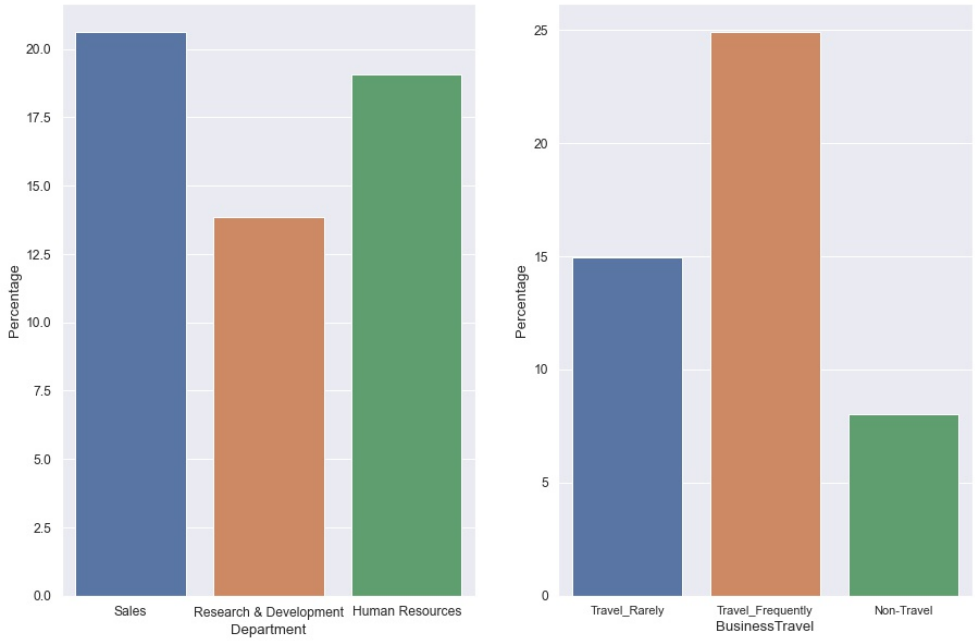

Figure 11-Percentage of leavers in terms of department and travel frequency

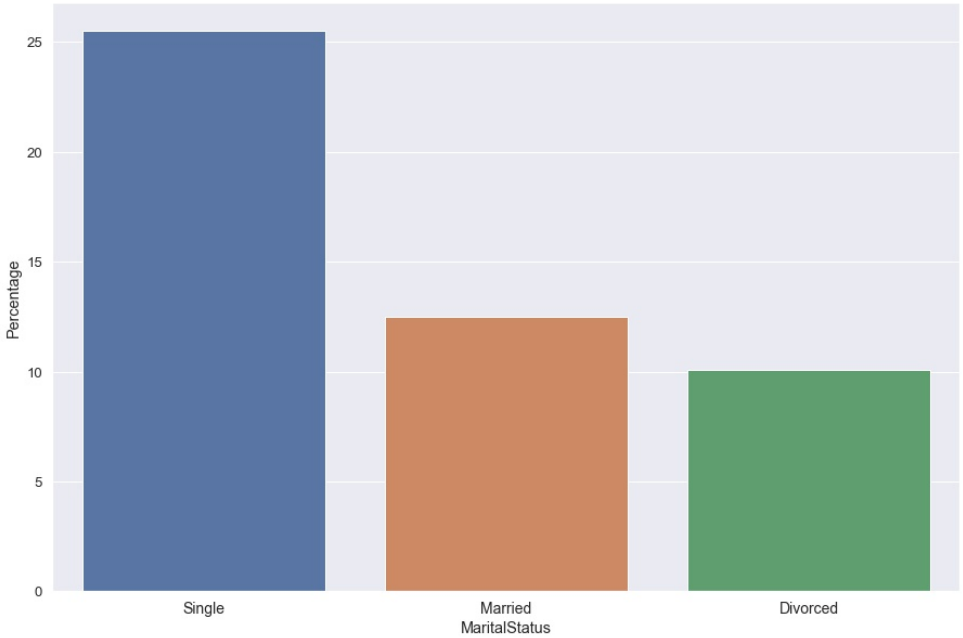

Figure 12-Percentage of leavers in terms of marital status

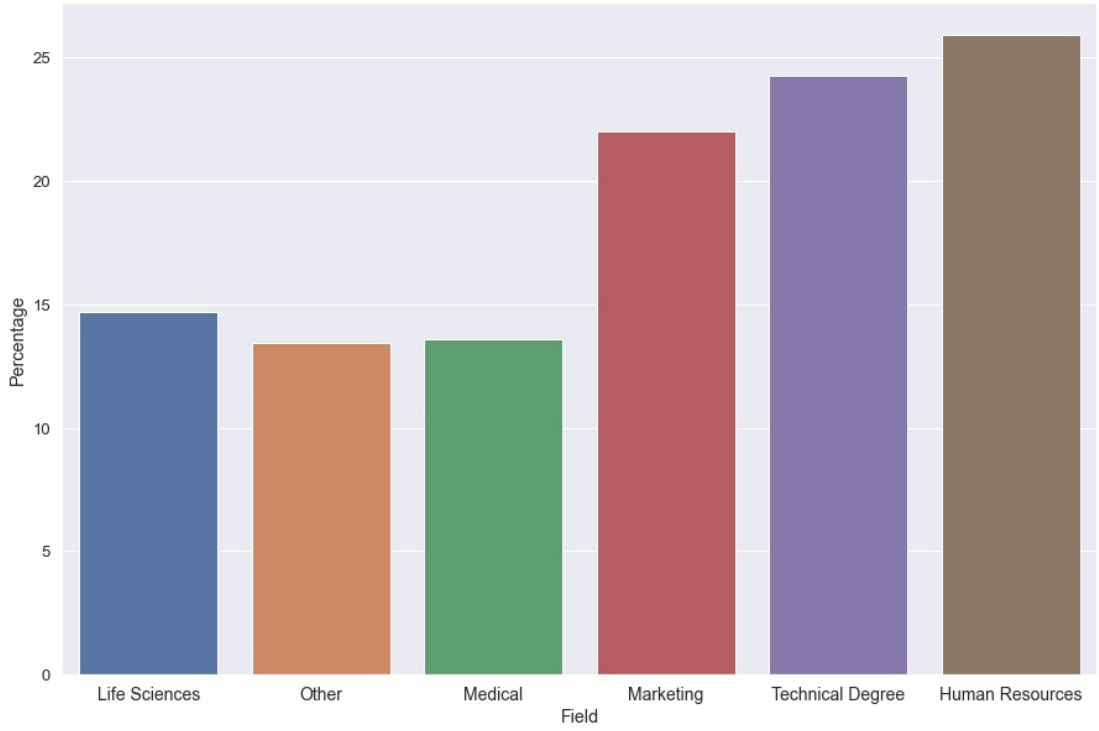

Figure 13-Percentage of leavers in terms of education field

Table 7-Categories Ranking

Rank	Income Status	Gender	Over time	Department	Business Travel	Marital Status	Education
1	Low: 18.94 %	Male: 17.00 %	Yes: 30.53%	Sales: 20.63 %	Travel Frequently: 24.91%	Single: 25.53%	Human Resources: 25.93 %
2	High: 10.55 %	Female: 14.80%	No: 10.44%	Human Resources: 19.04 %	Travel Rarely: 14.96 %	Married: 12.48%	Technical degree: 24.24 %
3				Research & Development: 13.84 %	Non Travel: 8%	Divorced: 10.09 %	Marketing: 22.01 %
4							Medical: 13.58%
5							Other: 13.41%
6							Life Sciences: 14.69%

We notice several things from the table:

1- Employees who have low income comparing to the average employee salary tend to leave the company due to the financial instability

2- Employees who work overtime represent a big portion of quitters (30.53%) comparing to current employees (10.44%)

3- Sales department represent the biggest portion of ex-employees. This could be due to the demand that doesn't have any particular pattern.

4- Employees who travel frequently are more likely to quit the organization for stability purposes.

5- The HR department has the biggest proportion of leavers.

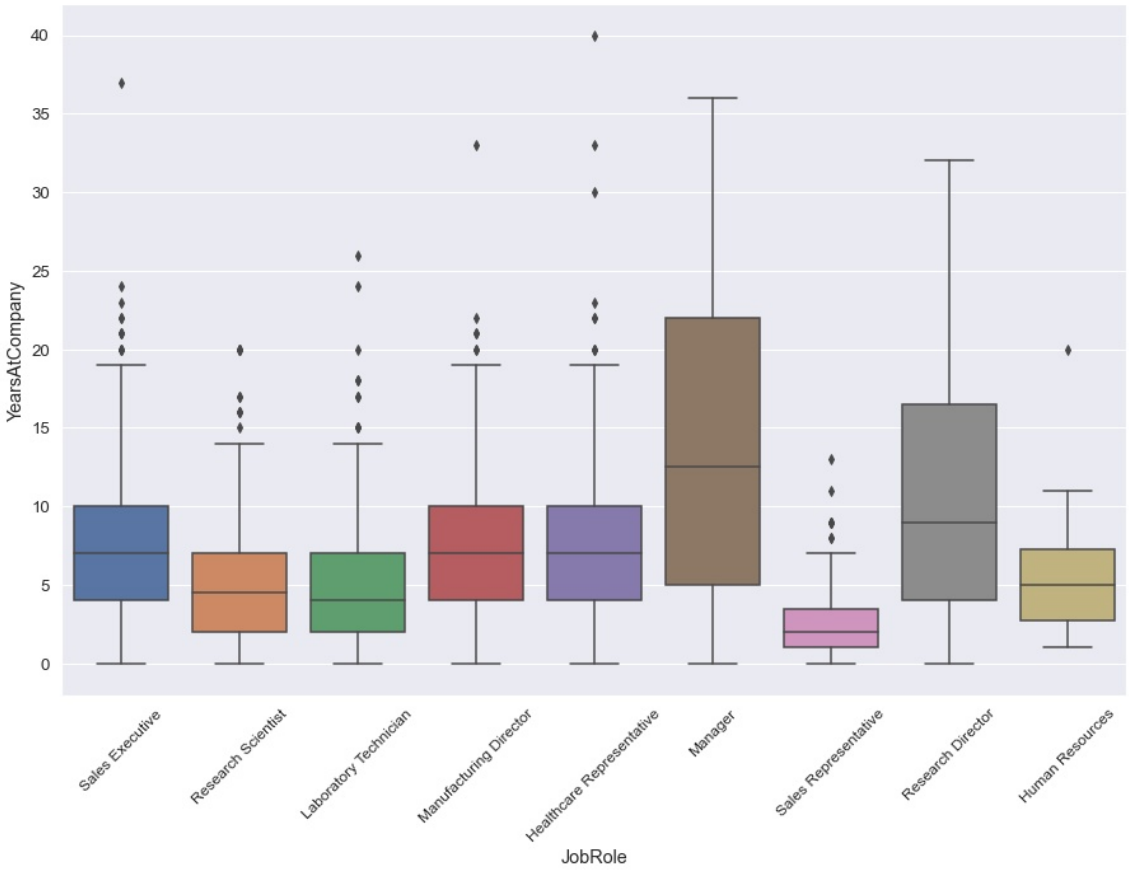

Figure 14-Boxplots for each job role

The average number of years at a company differs from a job role to the other. To become a manager, an employee should have stayed several years in the company. However, being a sales representative doesn't require many years.

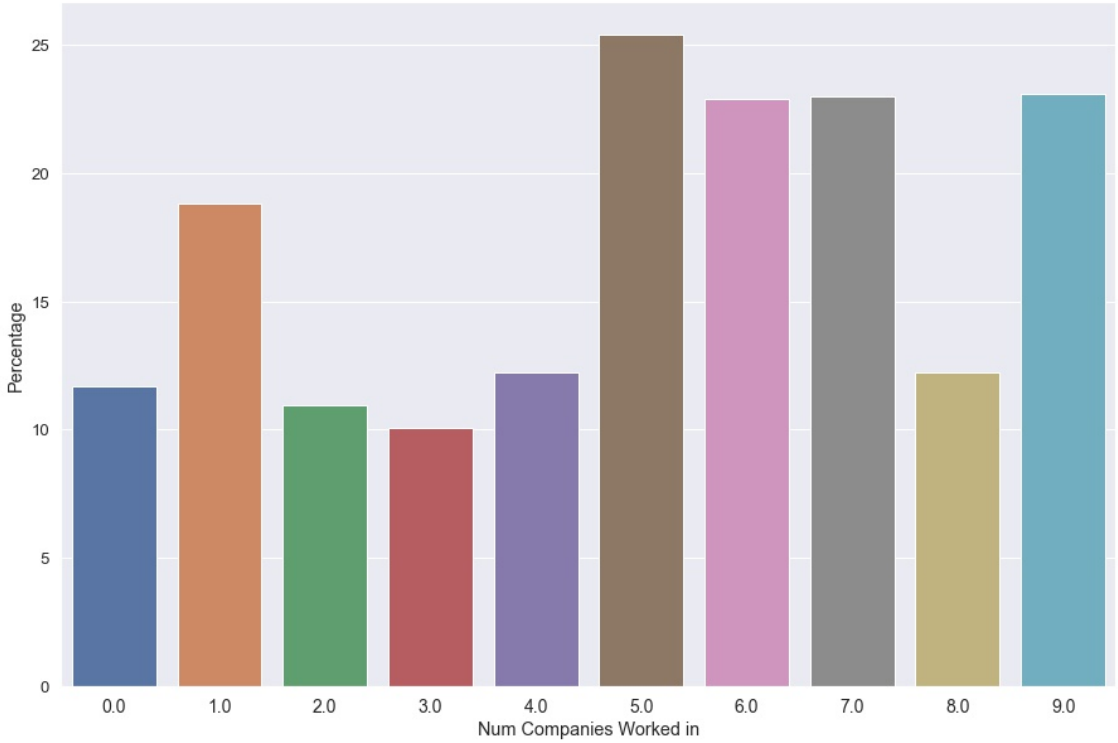

Figure 15-Percentage of leavers in terms of number of companies worked in

We observe that the highest percentages consist of employees that worked for more than 5 companies. This is due to the instability of the employee on a personal level since he/she keeps on switching companies.

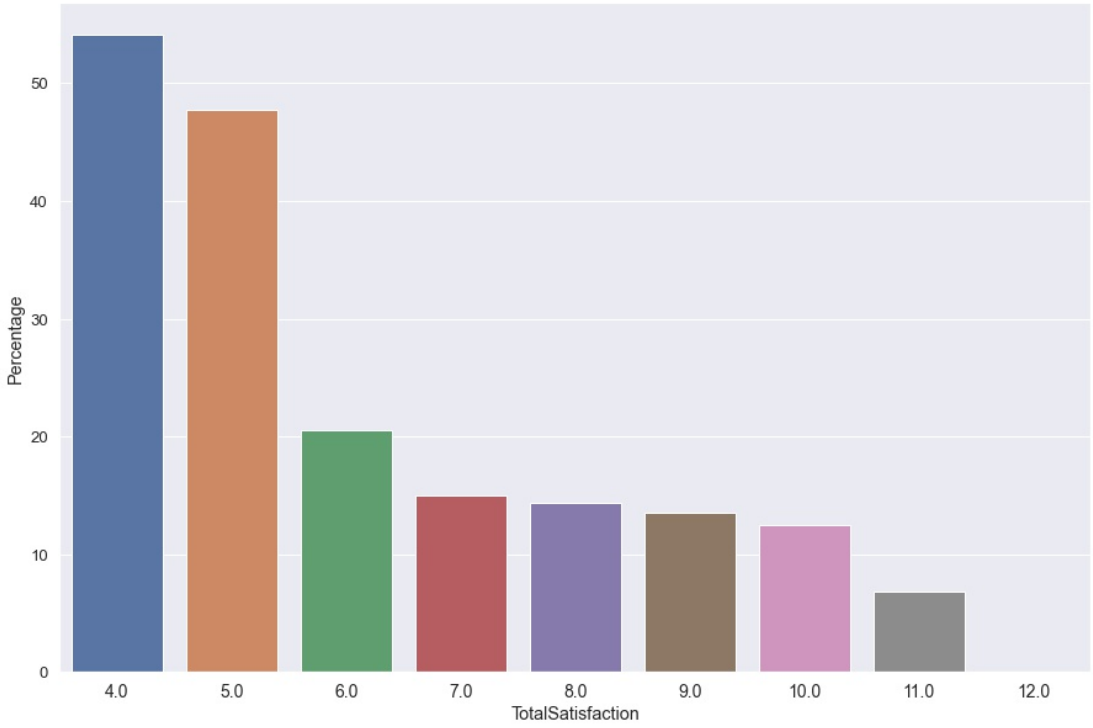

Figure 16-Percentage of leavers in terms of total satisfaction

The total satisfaction is the sum of environment satisfaction, job involvement, job satisfaction, performance rating, relationship satisfaction, and work-life balance. The highest percentage of leavers is for low values of the total satisfaction (Scores 4 and 5). Employees who leave their job aren't satisfied in their current position.

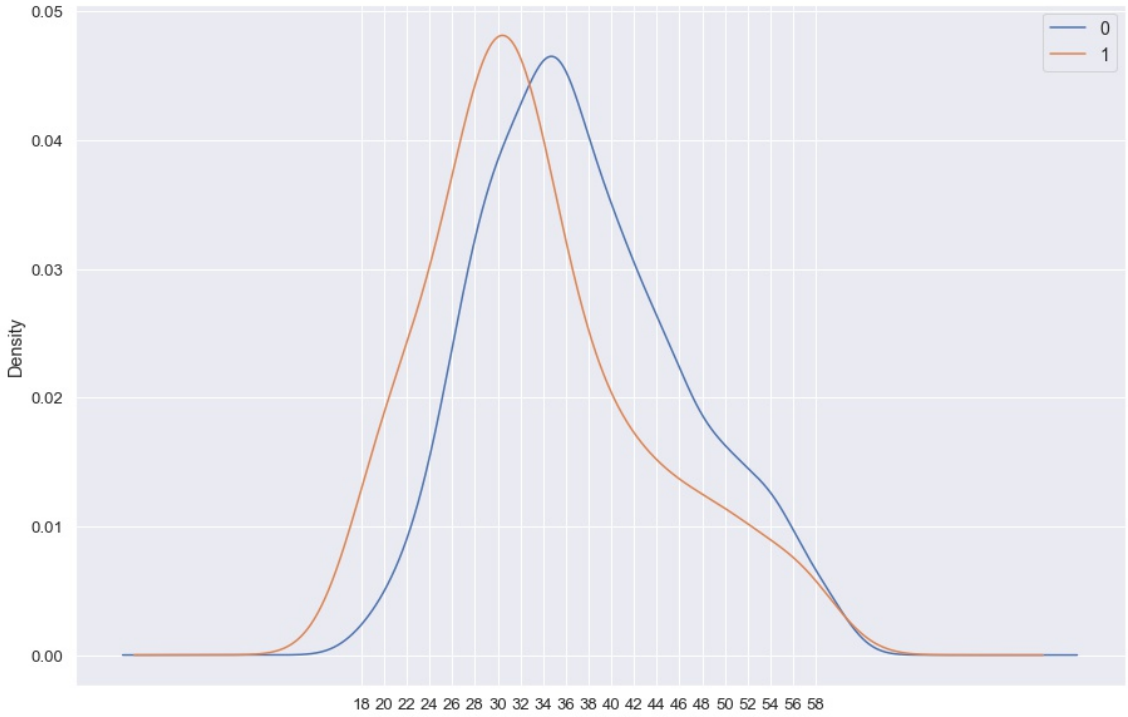

Figure 17-Age distribution

After doing computations, we conclude that the current employees' average age is 37.6 years old with a standard deviation of 8.9 However, the expected value of the ex-employees is 33.6 years old with a standard deviation of 9.7.

The average age of an employee leaving the company is much lower than the average age of a current employee.

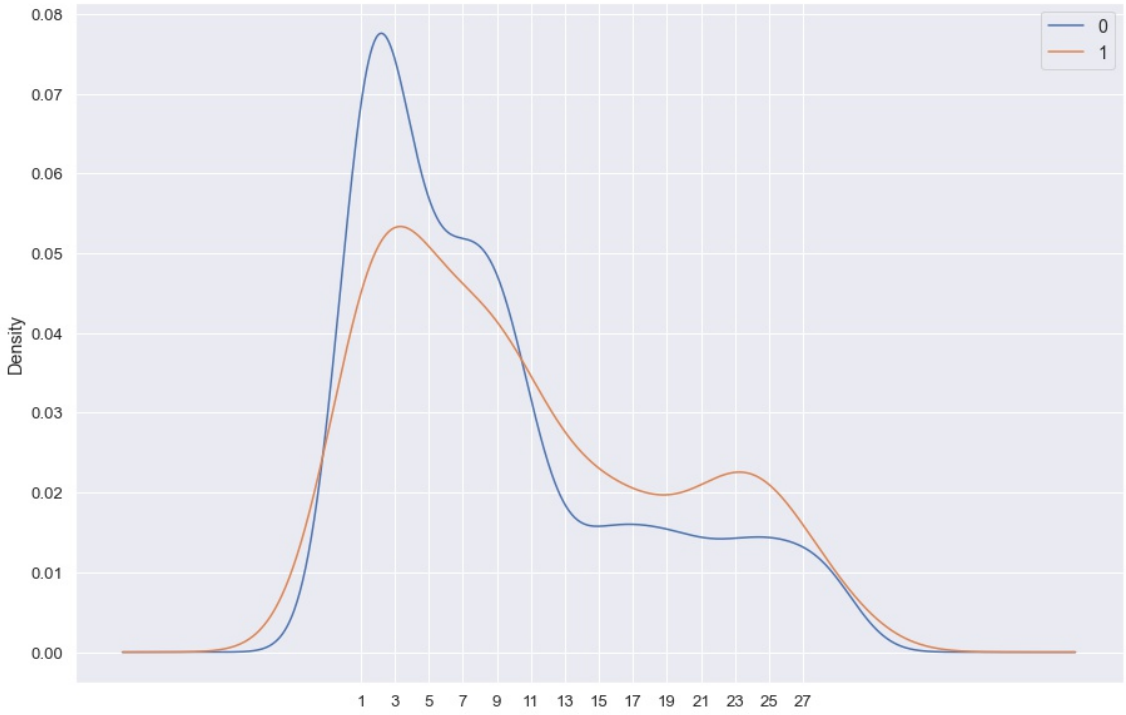

Figure 18-Home-work distance distribution

The average distance for active employees is 8.92 miles and for ex-employees 10.63 miles.

Chi-square is used to determine the dependence between two categorical variables. We need to do contingency tables to calculate the t-values and p-values of the Chi-square distribution. Contingency tables serve to visualize the frequency distribution of the variables. Chi-squared values depend on the observed (O) and the expected values (E):

$$X^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Table 8-Contingency tables examples

Attrition	0	1
Department		
Human Resources	0.809524	0.190476
Research & Development	0.861602	0.138398
Sales	0.793722	0.206278
All	0.838776	0.161224

Attrition	0	1
Gender		
Female	0.852041	0.147959
Male	0.829932	0.170068
All	0.838776	0.161224

Attrition	0	1
JobLevel		
1	0.736648	0.263352
2	0.902622	0.097378
3	0.853211	0.146789
4	0.952830	0.047170
5	0.927536	0.072464
All	0.838776	0.161224

Attrition	0	1
TotalSatisfaction		
4	0.458333	0.541667
5	0.522388	0.477612
6	0.794521	0.205479
7	0.849817	0.150183
8	0.856698	0.143302
9	0.864583	0.135417
10	0.875000	0.125000
11	0.931624	0.068376
12	1.000000	0.000000
All	0.838776	0.161224

Table 9-Chi-square outputs

Rank	Features	T-value	P-value
1	Over time	89.04	9.61×10^{-20}
2	Job Role	86.19	9.44×10^{-15}
3	Marital Status	46.16	5.23×10^{-10}
4	Business Travel	24.18	2.29×10^{-5}
5	Income Status	17.047	0.00019
6	Department	10.80	0.013
7	Education Field	16.025	0.014
8	Gender	1.28	0.53

So according to the Chi-square distribution, the most correlated feature to attrition is over time since it has the lowest p-value. All the categorical features listed above are statistically significant except for gender since its p-value is greater than 0.05.

B. Analytical results

Feature reduction is essential when working with many attributes. We notice that the IBM dataset contains 6 attributes having the same level of importance. For this reason, we can remove 6 variables (Environment Satisfaction, Job Involvement, Job Satisfaction, Performance Rating, Relationship Satisfaction, and Work-Life Balance) by grouping them into one variable called total satisfaction. Also, we remove the unnecessary attributes such as Employee Count, Employee Number, Over 18 (all employees are over 18 years old), standard hours, daily rate, monthly rate, and hourly rates. The reason why the rates were removed from the IBM dataset is reducing the number of variables and keeping the important rate which consists of the salary. To make it more representative, an additional variable called Income Status compared the salary of each employee to the expected salary of all employees. It is split into two categories: low and high.

To make our results more accurate, we applied hyperparameter tuning in random forest and logistic regression which helps us choose the best parameters to focus on. We set several values for different parameters, the grid search purpose is to loop over each value to find out the best parameters for a better result.

Table 10-Hyperparameters

Method	**Parameters**
Decision Tree	Criterion: Gini Class weight: Balanced
Random Forest	N_estimators: 50, 200, 300, 400 Min_samples_split: 2, 5, 10

	Bootstrap: True, False Class_weight: Balanced
AdaBoost	N_estimators: 50, 100, 200, 300, 400 Criterion: Gini, Entropy Splitter: Best, random Class_weight: Balanced
SVM	Kernel type: rbf Gamma: Auto Cache size: 200 MB Regularization Cost: 1 Probability: True
KNN	Number of neighbors K: 5
Logistic Regression	C: 0.001, 0.01, 0.1, 1, 10, 100 Solver: liblinear Class_weight: Balanced
Multi-Layer Perceptron	Solver: lbfgs Hidden Layer Sizes = 70 x 25 Activation = Logistic

We apply the label encoder function on the non-numeric columns. Label encoding is used for Random Forest, Decision Tree, Naive Bayes, Logistic Regression, and AdaBoost algorithms. We obtain the following summary statistics:

Table 11-Summary statistics

	Attrition	Age	BusinessTravel	Department	DistanceFromHome	Education	EducationField	Gender	JobLevel	JobRole
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	0.161224	18.923810	1.607483	1.260544	8.192517	1.912925	2.247619	0.600000	1.063946	4.458503
std	0.367863	9.135373	0.665455	0.527792	8.106864	1.024165	1.331369	0.490065	1.106940	2.461821
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	12.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	2.000000
50%	0.000000	18.000000	2.000000	1.000000	6.000000	2.000000	2.000000	1.000000	1.000000	5.000000
75%	0.000000	25.000000	2.000000	2.000000	13.000000	3.000000	3.000000	1.000000	2.000000	7.000000
max	1.000000	42.000000	2.000000	2.000000	28.000000	4.000000	5.000000	1.000000	4.000000	8.000000

TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	IncomeStatus	TotalSatisfaction
1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
11.278231	2.799320	6.991156	4.229252	2.187755	4.123129	0.664626	4.180272
7.775842	1.289271	6.053027	3.623137	3.222430	3.568136	0.472282	1.689520
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6.000000	2.000000	3.000000	2.000000	0.000000	2.000000	0.000000	3.000000
10.000000	3.000000	5.000000	3.000000	1.000000	3.000000	1.000000	4.000000
15.000000	3.000000	9.000000	7.000000	3.000000	7.000000	1.000000	5.000000
39.000000	6.000000	36.000000	18.000000	15.000000	17.000000	1.000000	8.000000

We use another type of data preprocessing for SVM, KNN, and ANN which is the one-hot encoding. As a rule of thumb, we apply one-hot encoding when the categorical features are not ordinal and label encoding when they are ordinal.

As we see in the summary statistics, the number of employees who quitted the company is much less than the ones who stayed. Since the dependent variable shows more values of “Yes” than “No”, the accuracy will be high even if the prediction is not accurate. For this reason, we will focus on the following evaluation metrics: sensitivity, recall, F1-measure, and ROC. Furthermore, we implemented a technique called SMOTE (Synthetic Minority Over-sampling Technique). This approach consists of over-sampling the minority class by creating artificial records rather than replicating them to avoid overfitting. It will force the decision region to become more generic and avoids particular cases (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). One of the studies showed a 5% approximate increase in accuracy after having a balanced data by

applying different machine learning algorithms such as Decision Tree, SVM, KNN, and Random Forest (Al Helal, Haydar, & Mostafa, 2016).

K-fold Cross-validation was also used for evaluation. The procedure consists of shuffling the data randomly, divide them into K groups. The model is evaluated on the k-1 folds through the first fold serving as a validation set. Then, K iterations are performed; the performance will be the average of each performance in each fold (James, Witten, Hastie, & Tibshirani, 2013).

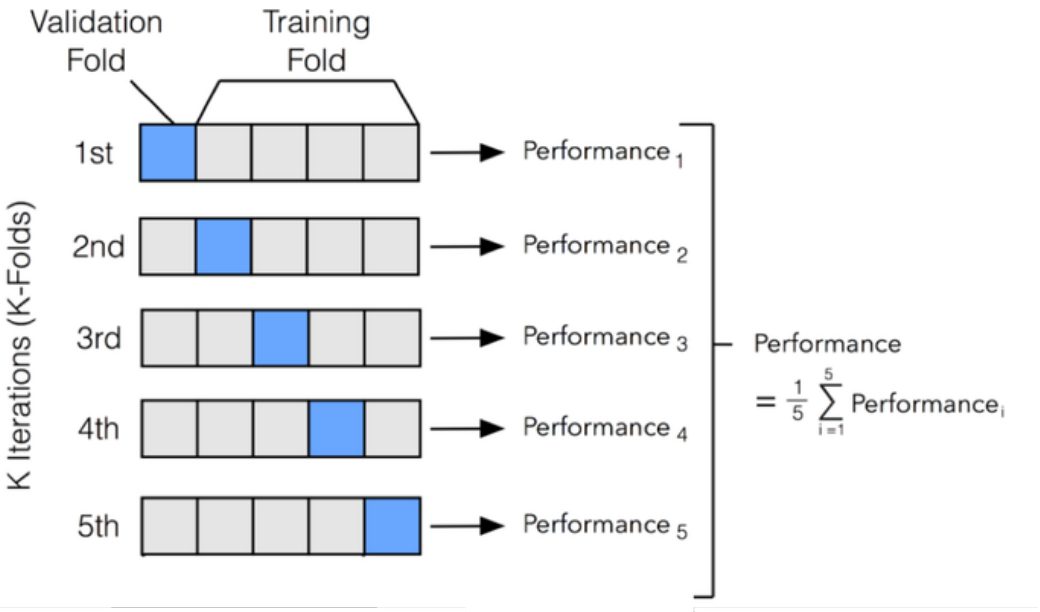

Figure 19-K-folds cross validation

It can be used with the SMOTE method (Santos, Soares, Abreu, Araujo, & Santos, 2018).

We standardize using the standard scaler in Python from the Sklearn library. It removes the mean from each x variable and divides it by the standard deviation through the following formula:

$$Z = \frac{(x - \mu)}{\sigma}$$

Then, we split the data into training and test set. We took 75 % training data and 25% test data. So we get 1102 instances of training data and 368 instances of the test set.

We used the oversampling technique where it randomly selects records from the minority class and duplicates them. We didn't use the under-sampling method where it removes instances from the majority class because we already have a small dataset. So the training set is formed of 1846 records instead of 1102 which is equivalent to an increase of 67.5%.

Subsequently, we apply machine learning algorithms. To get better results, grid-search was performed. The random forest over different values of n_estimators to get the best accurate results. Then, we train the model accordingly by using the training data from the SMOTE technique.

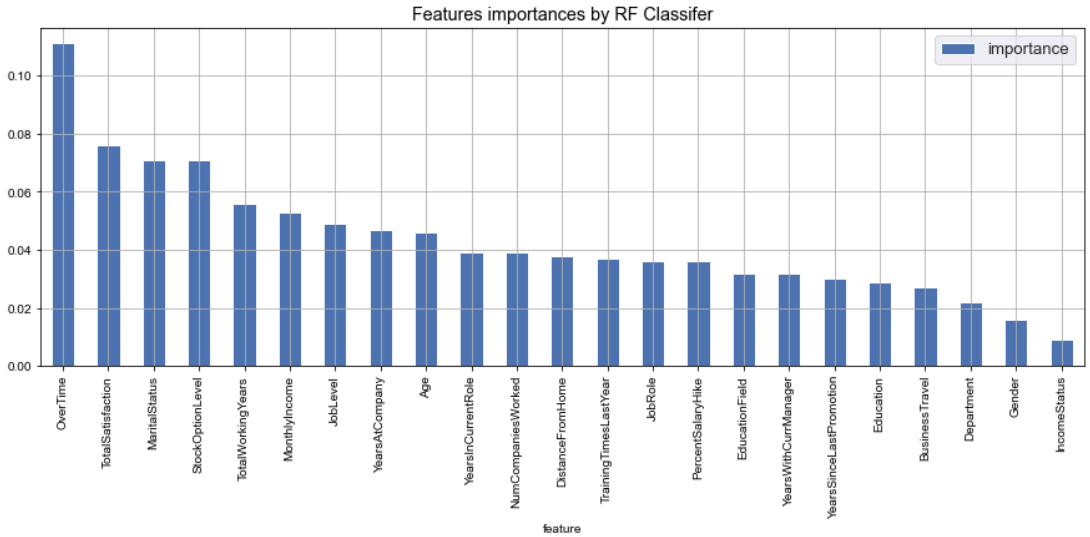

Figure 20-Features importance by Random Forest classifier

We notice that the most important features that lead to attrition are organizational factors rather than personal factors such as overtime, the total satisfaction, stock options level, total working years, and monthly income. An important personal factor that leads to attrition is marital status.

Table 12-Features importance values by Random Forest Classifier

feature	importance
OverTime	0.111
TotalSatisfaction	0.076
MaritalStatus	0.071
StockOptionLevel	0.071
TotalWorkingYears	0.056
MonthlyIncome	0.053
JobLevel	0.049
YearsAtCompany	0.047
Age	0.046

By applying several machine learning algorithms, we obtain the following results:

Algorithm	Accuracy	Precision	Recall	ROC
Random Forest	0.864	0.696	0.271	0.823
Decision Tree	0.774	0.329	0.390	0.619
Naïve Bayes	0.625	0.263	0.746	0.747
Logistic Regression	0.780	0.393	0.610	0.822
AdaBoost	0.780	0.311	0.322	0.593
SVM	0.848	0.528	0.475	0.807
KNN	0.685	0.276	0.593	0.711
ANN	0.829	0.462	0.407	0.782

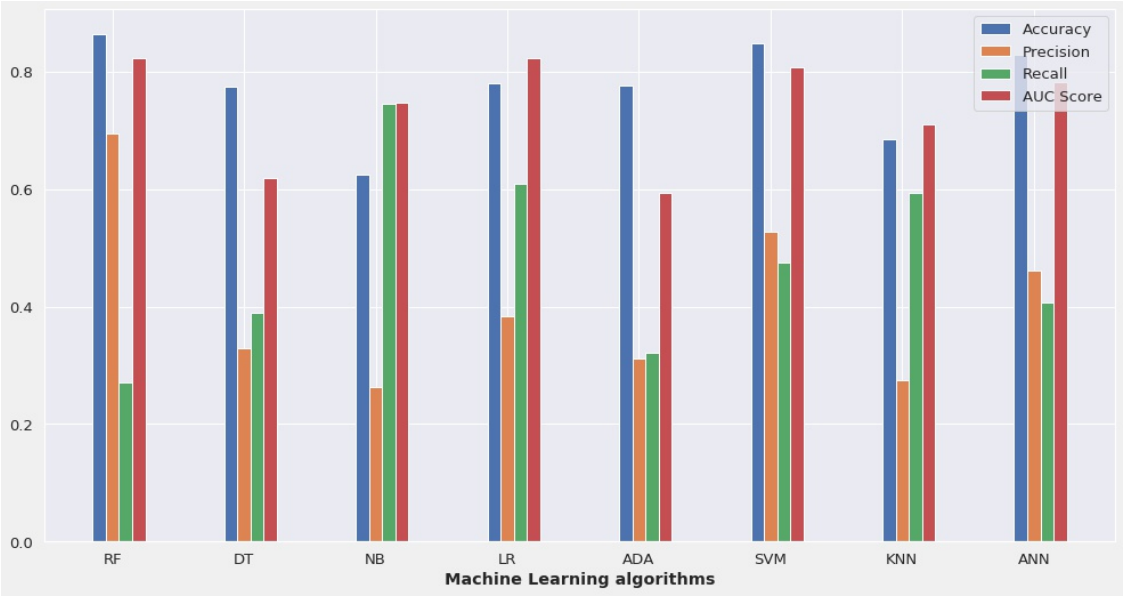

Figure 21-Evaluation metrics for different algorithms

The ROC curve is represented by the following picture:

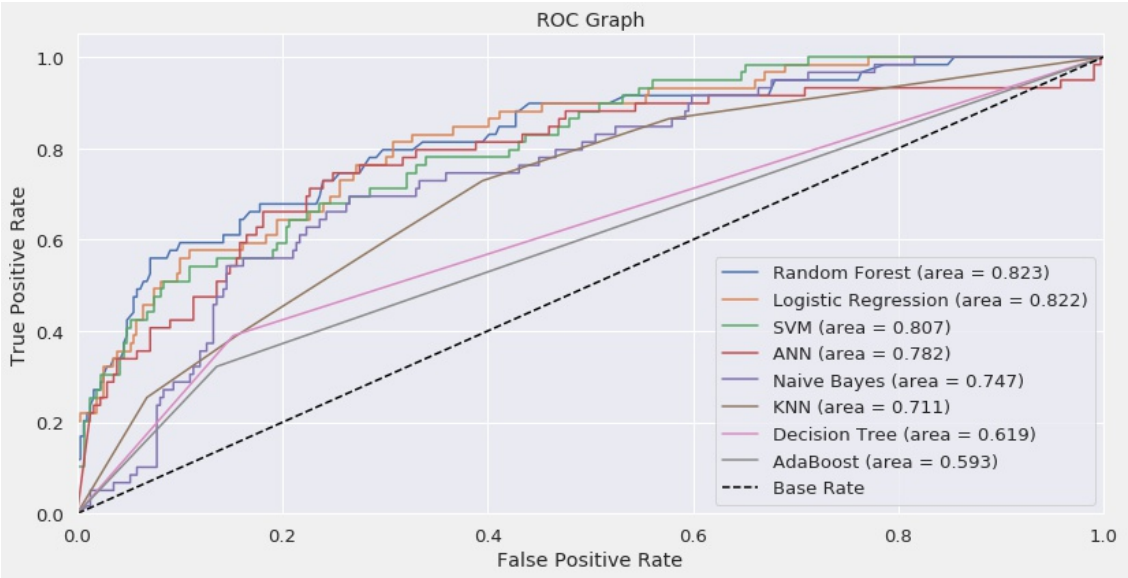

Figure 22-ROC graph

Based on our results, accuracy, precision, and AUC are the highest for the random forest classifier. It has an area under the curve of 0.823.

After filtering the data by taking into account the ex-employees, a linear regression was performed to know the most correlated features related to the total number of years spent by the employee in the company. A regression model helps identifies a mathematical formula with the Y variable which is the number of years at the company.

A regression model was applied to understand the correlation between the features with the number of years at the company for ex-employees.

Table 13-Regression results

OLS Regression Results

Dep. Variable:	YearsAtCompany	R-squared (uncentered):	0.683
Model:	OLS	Adj. R-squared (uncentered):	0.671
Method:	Least Squares	F-statistic:	56.10
Date:	Sun, 11 Apr 2021	Prob (F-statistic):	3.52e-42
Time:	17:31:57	Log-Likelihood:	-555.40
No. Observations:	189	AIC:	1125.
Df Residuals:	182	BIC:	1147.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Age	0.2062	0.042	4.855	0.000	0.122	0.290
DistanceFromHome	-0.0963	0.039	-2.470	0.014	-0.173	-0.019
MonthlyIncome	0.0075	0.001	7.269	0.000	0.005	0.010
NumCompaniesWorked	-0.4905	0.129	-3.803	0.000	-0.745	-0.236
PercentSalaryHike	-0.0614	0.093	-0.660	0.510	-0.245	0.122
StockOptionLevel	-0.0509	0.386	-0.132	0.895	-0.812	0.711
TotalSatisfaction	0.2423	0.147	1.650	0.101	-0.047	0.532

Omnibus:	40.936	Durbin-Watson:	2.045
Prob(Omnibus):	0.000	Jarque-Bera (JB):	96.586
Skew:	0.951	Prob(JB):	1.06e-21
Kurtosis:	5.940	Cond. No.	739.

We chose the features that have a p-value less than 0.05 which shows a strong dependence on the total number of years at the company.

We notice that the most correlated features related to employee attrition are listed below:

1- Age (coefficient: 0.2062, positive correlation)

2- Monthly Income (coefficient: 0.0075, positive correlation)

3- Number companies worked (-0.4905, negative correlation)

4- Distance from home (-0.0963, negative correlation)

We obtain interesting results. We notice that age and monthly income are strongly correlated to the number of years an employee staying at the company. The higher the income is, the more the employee will commit to the company. We also notice a negative correlation between the number of companies an employee worked in with the number of years at the company. If an employee worked in many companies before, that means that this employee is not stable in his/her job, that's why he/she is more likely to leave the company. Also, distance from home to work has a negative correlation with the number of years an employee stayed at the company. Employees who live farther away from work are more likely to quit their job since the long distance to work might be exhausting. So, the employee will search for a nearer job location.

The errors obtained from the regression model were low:

- Mean absolute error: 3.06
- Mean squared error: 17.19
- Root mean squared error: 4.15

CHAPTER V

CONCLUSION

We presented in the paper several approaches to understand the reasons why employees quit their job, from descriptive analysis to predictive analysis. Our approach of evaluating the models by balancing the data with hyperparameter tuning was robust and provided accurate results. We found out that the random forest classifier was the best one in terms of accuracy, precision, and area under the curve. A regression model was built to predict 'years at the company' for those who left (attrition=1). We used the features as predictor variables and fit a linear model to understand which attributes are positively correlated and which are negatively correlated with the target variable.

After applying several machine learning algorithms and deep learning, we found out that the random forest classifier is the best one. After balancing our data, our results outperformed the results from the literature review.

Table 14-Comparative Results

Janusz et al., 2018	**Dolatabadi & Keynia, 2017**	**Pessach et al., 2020**	Our Work
AUC of RF = 0.641	RF accuracy 79 %	AUC of RF = 0.71	AUC of RF = 0.823

After summarizing the descriptive and analytical part, the most important features are overtime, the total satisfaction, stock options level, distance from home, marital status, and monthly income.

Recommendations:

- HR should consider the following attributes when an employee is working overtime, the total satisfaction of the employee in the company that takes into account environment satisfaction, job satisfaction, relationships satisfaction, job involvement, performance rating, and work-life balance.
- Attrition tends to be higher in the younger age groups; our data showed that those who left the company were around 4 years younger on average than those who have not. Overtime should be monitored since it is the most important attribute in the descriptive, statistics, and predictive analysis.
- HR should take into consideration that employees' monthly income might lead to employee attrition if the employee's salary is below average. Then, stock options should be encouraged in the forms of contracts that will allow the employee to have shares in the company's shares. The shares could be set for an amount of time.
- HR should evaluate the employee's CV. If he/she keeps on switching companies, then this employee is more likely to leave the company.

Additional Retention strategies (based on common practice):

- Keeping on evaluating employees' mental health throughout the year by checking if the person is satisfied and involved in their job, satisfied with the environment, colleagues, manager, and having a good work-life balance.
- Giving Yoga sessions for the employees. Many companies are providing yoga sessions for their employees that help them relax and boost their productivity.

- If the company can’t provide their employees with stock options level, an alternative would be to offer benefits packages, challenging salaries, and bonuses.
- Being transparent and clear to solve every problem and avoid any kind of tension between the manager and the employees.
- Most importantly, a manager should be a great leader possessing a high emotional intelligence. His/her role would be to be a motivational leader whose employees will look up to. If the manager isn’t passionate and doesn’t know how to lead a team, probably all the team will fail and will quit their job.

CHAPTER VI

LIMITATIONS AND FUTURE WORK

A. Limitations

The datasets consist of past information and employee attrition but don't contain any accurate time when the employee left the company. The turnover problem can be biased since it doesn't take into account the current employee job. An employee might leave the company because of a specific or unpredictable event that can occur in the present. Another limitation of this study is the lack of information in the dataset. An additional parameter suggesting the employee activity after quitting the job may lead to more accurate results regarding the attrition. If the employee leaves a company to complete his studies or to have an additional degree has a different impact than an employee who quitted his/her job to get hired in another firm.

B. Future Work

Future work can consist of creating our dataset and choosing our features. Besides, the data would incorporate time variants and the exact date of attrition. The dataset would be much larger, thus, having more accurate results. The number of features can be increased in a way having clusters that include the external environment and structural factors. The external environment factor reveals employee voluntary attrition. It can suggest job opportunities outside the firm and different job conditions. Another cluster that can be implemented is structural factors that reveal the employees' networking skills and the ability to cope with their colleagues and bosses.

BIBLIOGRAPHY

Abbott, J. (2003). Does employee satisfaction matter? A study to determine whether low employee morale affects customer satisfaction and profits in the business-to-business sector. *Journal of Communication management, 7*(4), 333-339.

Al Helal, M., Haydar, M. S., & Mostafa, S. A. M. (2016). *Algorithms efficiency measurement on imbalanced data using geometric mean and cross validation.* Paper presented at the 2016 International Workshop on Computational Intelligence (IWCI).

Anand, V. V., Saravanasudhan, R., & Vijesh, R. (2012). *Employee attrition-A pragmatic study with reference to BPO Industry.* Paper presented at the IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012).

Bach, M. P., Simic, N., & Merkac, M. (2013). Forecasting Employees' Success at Work in Banking: Could Psychological Testing Be Used as the Crystal Ball? *Managing global transitions, 11*(3), 283-299. Retrieved from http://aub.summon.serialssolutions.com/#!/search?ho=t&fvf=ContentType,Book%20Review,t&l=en&q=(Forecasting%20Employees%E2%80%99%20Success%20at%20Work%20in%20Banking:%20Could%20Psychological%20Testing%20Be%20Used%20as%20the%20Crystal%20Ball%3F)

Baig, M. M., Awais, M. M., & El-Alfy, E.-S. M. (2017). AdaBoost-based artificial neural network learning. Neurocomputing, 248, 120-126.

Bronshtein, A. (2019). A quick introduction TO k-nearest Neighbors ALGORITHM. Retrieved April 04, 2021, from https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

Cai, X., Shang, J., Jin, Z., Liu, F., Qiang, B., Xie, W., & Zhao, L. (2020). DBGE: Employee Turnover Prediction based on Dynamic Bipartite Graph Embedding. *IEEE Access, 8*, 10390-10402.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In (pp. 853-867). Boston, MA: Springer US.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chien, C.-F., & Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications, 34*(1), 280-290.

Cotton, J. L., & Tuttle, J. M. (1986). Employee turnover: A meta-analysis and review with implications for research. *Academy of management Review, 11*(1), 55-70.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management, 45*(3), 164-174.

Dolatabadi, S. H., & Keynia, F. (2017). *Designing of customer and employee churn prediction model based on data mining method and neural predictor.* Paper presented at the 2017 2nd International Conference on Computer and Communication Systems (ICCCS).

Erik van Vulpen. (2020). (AIHR). How to calculate employee turnover rate. Retrieved February 17, 2021, from https://www.analyticsinhr.com/blog/how-to-calculate-employee-turnover-rate/#:~:text=How%20to%20calculate%20annual%20turnover,%3D%200.05%2C%20or%205%25.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1): Springer series in statistics New York.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., . . . Sriram, S. (2016). Modeling Customer Lifetime Value. *Journal of service research : JSR, 9*(2), 139-155. doi:10.1177/1094670506293810

Holt, D. T., Armenakis, A. A., Feild, H. S., & Harris, S. G. (2007). Readiness for organizational change: The systematic development of a scale. *The Journal of applied behavioral science, 43*(2), 232-255.

Huang, Y., Huang, B., & Kechadi, M. (2010). *A new filter feature selection approach for customer churn prediction in telecommunications.* Paper presented at the 2010 IEEE International Conference on Industrial Engineering and Engineering Management.

(Vol. 112): Springer.

Jain, R., & Nayyar, A. (2018). *Predicting employee attrition using xgboost machine learning approach.* Paper presented at the 2018 International Conference on System Modeling & Advancement in Research Trends (SMART).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*

Janusz, A., Stawicki, S., Drewniak, M., Ciebiera, K., Ślęzak, D., & Stencel, K. (2018). *How to Match Jobs and Candidates-A Recruitment Support System Based on Feature Engineering and Advanced Analytics.* Paper presented at the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems.

Judge, T. A., Cable, D. M., Boudreau, J. W., & Bretz JR., R. D. (1995). An Empirical Investigation of the Predictors of Executive Career Success. *Personnel Psychology, 48*(3), 485-519. doi:10.1111/j.1744-6570.1995.tb01767.x

Mathan, K., Kumar, P. M., Panchatcharam, P., Manogaran, G., & Varadharajan, R. (2018). A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Design automation for embedded systems, 22(3), 225-242.

Mitkees, I. M. M., Badr, S. M., & ElSeddawy, A. I. B. (2017). *Customer churn prediction model using data mining techniques*.

Nappinnai, M., & Premavathy, N. (2013). Employee attrition and retention in a global competitive scenario. *International Journal of Research in Business Management (IMPACT: IJRBM), 1*(6), 11-14.

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 113290.

RayI, S. (2020, October 12). Learn naive BAYES Algorithm: Naive Bayes Classifier examples. Retrieved April 5, 2021, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Rothwell, W., Alexander, J., Bernhard, M., & Books24x, I. (2008). Cases in Government Succession Planning: Action-oriented Strategies for Public Sector Human Capital Management, Workforce Planning, Succession Planning and Talent Management. Amerherst: HRD Press.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. *ieee ComputatioNal iNtelligeNCe magaziNe, 13*(4), 59-76.

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications, 38*(3), 1999-2006.

Shah, N., Irani, Z., & Sharif, A. M. (2017). Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors. *Journal of Business Research, 70*, 366-378.

Singh, M., Varshney, K. R., Wang, J., Mojsilovic, A., Gill, A. R., Faur, P. I., & Ezry, R. (2012). *An analytics approach for proactively combating voluntary attrition of employees.* Paper presented at the 2012 IEEE 12th International Conference on Data Mining Workshops.

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017). *Evaluation of machine learning models for employee churn prediction.* Paper presented at the 2017 International Conference on Inventive Computing and Informatics (ICICI).

Sivaram, N., & Ramar, K. (2010). Applicability of Clustering and Classification Algorithms for Recruitment Data Mining in International Journal of Computer Applications (0975–8887) Volume 4–No. 5. In: July.

Soni, U., Singh, N., Swami, Y., & Deshwal, P. (2018). *A Comparison Study between ANN and ANFIS for the Prediction of Employee Turnover in an Organization.* Paper presented at the 2018 International Conference on Computing, Power and Communication Technologies (GUCON).

Verbeke, W., Hur, J., Martens, D., Dejaeger, K., & Baesens, B. (2012). *New insights into churn prediction in the telecommunication sector: a profit driven data mining approach*.

Wills, S., Underwood, C. J., & Barrett, P. M. (2021). Learning to see the wood for the trees: machine learning, decision trees, and the classification of isolated theropod teeth. Palaeontology, 64(1), 75-99.

Yadav, S., Jain, A., & Singh, D. (2018). *Early Prediction of Employee Attrition using Data Mining Techniques.* Paper presented at the 2018 IEEE 8th International Advance Computing Conference (IACC).

Yan, M., & Zhou, Z. (2010). *A SE Model on Panic of Employee Turnover.* Paper presented at the 2010 International Conference on Management and Service Science.

Yiğit, İ. O., & Shourabizadeh, H. (2017). *An approach for predicting employee churn by using data mining.* Paper presented at the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP).

Zhang, P.-B., & Yang, Z.-X. (2016). A novel adaboost framework with robust threshold and structural optimization. IEEE transactions on cybernetics, 48(1), 64-76.

Zhang, Y., Yan, G., & He, S. (2016). Optical Fiber Spectrometer based on Smartphone Platform for Refractive Index Sensing Application. Paper presented at the Asia Communications and Photonics Conference.

Zhu, Q., Shang, J., Cai, X., Jiang, L., Liu, F., & Qiang, B. (2019). *CoxRF: Employee turnover prediction based on survival analysis.* Paper presented at the 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).